Essays on model averaging and political economics

Wang, W.

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Essays on model averaging and political economics

Wendun Wang

November 8th, 2013

Essays on model averaging and political economics

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 8 november 2013 om 14.15 uur door

Wendun Wang

geboren op 17 december 1984 te Fuzhou, P.R.China.

Promotiecommissie:

Promotor:	prof. dr. Jan Magnus
Overige Leden:	prof. dr. Erwin Bulte
	prof. dr. Gerda Claeskens
	dr. Dmitry Danilov
	prof. dr. John Einmahl
	prof. dr. Arthur van Soest

Acknowledgements

This dissertation contains my four years of work in Tilburg University. These four years helped me develop my way of doing research and broad collaboration with faculties as well as Ph.D. student fellows. I have grown as a young researcher over these years with the help of many people, and it is my great pleasure to acknowledge their help and support.

The most important person in my Ph.D. study is my supervisor Prof. Jan R. Magnus, to whom I would like to express my immense gratitude. Prof. Magnus is far more than an academic supervisor to me, but also a wise guider of my life, amiable senior, humorous friend, et al. My life path is significantly and positively changed by Prof. Magnus. I still remember my first meet with Jan who wore a quite uniquely light yellow suit standing at the exit of Xiamen airport, and I was very impressed by your manners of a gentleman. If I had not meet you in Xiamen, I might now be boringly working as a mediocre office worker in some company in China. It is my great pleasure and honor to be your last Ph.D. student in Tilburg before you delivery your valedictory address. You have influenced me to a great extent with your rigorous attitude toward research and in-depth thinking. We have an old saying in China that

"*Giving a man a fish and feeding him for a day is not better than teaching him how to fish and feeding him for a lifetime*".

This is exactly what you do to me. For example, you often suggest me not to work too quickly to make mistakes, but more importantly be precise and correct. This calms me down, and let me sit there and work peacefully to have more useful thinking. By every chance, you teach me how to think and write in a logic, precise, and concise way, from which I benefit a lot. You gave me much freedom to develop my own research agenda and the flexibility to collaborate with different researchers. You also gave me plenty of precious advice for my future research agenda. Besides your incredible knowledge and academic experience, you also shared life stories with my wife and me, and gave us valuable advice of balancing between life and work. We are so moved by your warm wishes for our new-born daughter. I am deeply indebted to your support and help which go beyond the research.

I am indebted to my coauthors who made great contributions to the progress of the work contained in my dissertation. Xinyu Zhang, you are such a pleasant person to work with. When we brainstormed together, I was constantly amazed by your statistical skills and your attention to details. For the last chapter of my dissertation, my thanks also go to Ruixin Wang. As the coauthor of my last chapter, thank you so much for sharing your great research ideas, your diligent and efficient work. I have benefited a lot from daily discussion with you and I am impressed by your enthusiasm for economic research. I appreciate our collaboration and I hope that our collaboration will continue in the field of econometrics and political economics. I am also extremely grateful to coauthors of other papers work in progress. To Gerda Claeskens, thank you very much for your hospitality and enlightening discussion when we visit leuven; To Zongxin Qian and Jinghua Lei for inspiring discussion and great efforts on our work.

For this dissertation I would like to thank my reading committee members: Erwin Bulte, Gerda Claeskens, Dmitry Danilov, John Einmahl, and Arthur van Soest for their time, interest, and insightful comments. I have also learnt a lot from inspiring discussions with Pavel Čížek, Bertrand Melenberg, Patricia Prüfer and Ramon van den Akker. I appreciate seminar and conference participants for useful comments, especially in Erasmus University Rotterdam and Rijks University Groningen.

I will never forget the suggestions and support offered during my stressful job search period. I was greatly supported by my references: Jan Magnus, Frederic Vermeulen, and Erwin Bulte. I am thankful for the useful advice from the Tilburg Ph.D. job search committee: Johannes Binswanger, Otilia Boldea, and Meltem Daysal. John Einmahl, Arthur van Soest, and Ronald Mahieu also shared with me their helpful tips on my mock interview. I also received enormous help and encouragement from Fangfang Tan, Juanjuan Cai, and Zongxin Qian who were on the job market one or two years ago.

I am indebted to my many student colleagues for providing a stimulating and fun environment in which to learn and grow. I sincerely thank my friend Zongxin Qian. Although you have been very busy since you embarked on your career as assistant professor in China, we still managed to have regular communications and discussions about our research. These discussions enable me to have a closer look of my work. My special thanks also go to my office mate Serhan Sadikoglu, and former office mate Noelia Lobato. Sharing an office with you during these years was one of the pleasures of doing a Ph.D. It was good to have someone to share and discuss research with. Serhan, it is always fun to chat with you about the football games between top two Turkish football teams, the current affairs and traditions in China and Turkey. I wish you good luck on the job market next year!

The department of Econometrics in Tilburg University offers a great working environment for Ph.D. students. I would like to thank Korine, Anja, Heidi, and Lenie for their efficient work and warm wishes when my daughter was born. I also thank all secretaries from the secretariat in CentER Graduate School, especially Ank, Corine, Aafke, and Jasmijn for assisting me in many different ways. I would also like to thank Prof. Henk Norde, Prof. Carol Burgers, Servaas van Bilsen, Jochem de Bresser, Chao-Hui Koo, Sybren Huijink, and Marieke Quant for the nice teaching experience we shared together.

My time at Tilburg University was made enjoyable in large part due to the many friends that became a part of my life. The wonderful companionship of Te Bao, Juanjuan Cai, Zhenzhen Fan, Feng Fang, Zhongjiong Gan, Yufeng Huang, Masako Ikefuji, Hiroaki Sakamoto, Xue Jia, Kamlesh Kumar, Jinghua Lei, Hong Li, Zengyu Li, Kebin Ma, Ning Ma, Miao Nie, Geng Niu, Zihan Niu, Fangfang Tan, Keyan Wang, Yun Wang, Ran Xing, Ying Yang, Huaxiang Yin, Yuejuan Yu, Bin Zhou, and the badminton group ensures that I can only think back upon my days in Tilburg with feelings of fondness and reminisce. Thank you very much for the friendship!

I owe my immense gratitude to my family, especially mama, baba, and laolao for their unconditional love and encouragement. After I have my own child, I understand more your hard feeling of being away from your only child for so many years. Thank you so much for your respect and support about my decision. Lastly, I would like to express my deepest love to Kan Ji, my best friend, soul-mate, and wife. These past several years have not been an easy ride, both academically and personally. My life is made much happier by you, and you can always bring me the "fresh air" with your knowledge and a lot of fun everyday. We both learnt a lot about life and strengthened our commitment to each other.

Contents

Chapter 1

Introduction

Literally, 'economics' is a strict subset of 'econometrics', but their relation in practice is far more complicated than containing. It is also a non-trivial job to associate them in an appropriate way in a study, and I find it even more difficult to do this in a Ph.D. thesis, at least for me. Nevertheless, I find it equally interesting if we could apply appropriate econometric methods in finding the answer to an economic question, or develop a well-suited econometrics method to address an economic problem. This thesis tries to do these two things. In the first two chapters we try to develop general and hopefully appropriate econometric tools to address problems encountered by economists. In the next two chapters, we try to investigate the effect of some policies using appropriate econometric methods.

1.1. Estimation and prediction under uncertainty

Chapter 2 and 3 study estimation and prediction in an uncertain environment. Researchers and policy-makers are exposed to a world full of uncertainty, and thus their studies and decisions have to be made under various sources of uncertainty. Let me borrow a story from my supervisor. Suppose a ruler seeks advice on a specific parameter, say next year's inflation. He has twelve advisors, and each advisor provides an estimate. When all have left, the ruler has twelve estimates. In addition, he has an opinion about each advisor based on past experience and current performance. How does the ruler now obtain a single estimate? There are at least two possibilities. The ruler may think: Whom do I trust most? Whose advice do I think most reliable? Then, he takes the advice of his most trusted advisor. This is the first method. Alternatively, he may consider all advisors useful, but not to the same degree. Some are more experienced and more clever than others, so they get a higher weight. Then, the ruler computes a weighted average of the twelve estimates. This is the second method. While the second method appeals to common sense, econometric practice favors the first method. In econometric practice one typically first selects the 'best' model based on diagnostic tests (such as t-ratios, R^2, and various information criteria) and then computes estimates within this selected model. This is called 'pretesting'. There are many problems with this procedure. For example, the pretest estimator is 'kinked'; it has a discontinuity at one. This is not only mathematically undesirable but also intuitively: we exclude a variable if its estimated coefficient has a t-ratio less than 1.96; we include it if the t-ratio is larger than 1.96. Most importantly, model selection and estimation are completely separated—just like the ruler only listening to his most trusted advisor—so that uncertainty in the model selection is ignored when reporting properties of the estimates.

In these two chapters we use model averaging techniques to incorporate the uncertainty. In the first chapter, we consider estimating a regression model when there are multiple measurements for independent variables. In particular, when one specifies the equation to be estimated, one has to decide which concepts (say inflation) to include in the regression. In addition, one has to decide which measurements of these concepts to use (for example, CPI-based or PPI-based inflation). This gives rise to two levels of uncertainty: concepts (level 1) and measurements within each concept (level 2), and we call this the 'measurement problem'. The measurement problem raises at least three issues. First, different choices of measurements produce different estimates for the same concept, leading to ambiguity in explanation and policy implications. Second, multiple measurements typically cause multicollinearity if they are included in one regression model, so that the estimates for individual measurements are imprecise, and statistical inference on a concept based on these estimates can therefore be misleading. Third, including multiple measurements in one model can also cause a problem of dimensionality when the number of explanatory variables is close to or even exceeds the number of observations.

Although this problem has been realized in the growth literature by Brock et al. (2003), there are no completely satisfactory solutions. One conventional method is to use the robustness check or extreme bounds analysis (Leamer and Leonard, 1983). The main drawback of this method is that it produces different estimates for each concept, and requires too strong conditions to obtain robust results. Alternatively, one would often try many different concepts, and select the most appealing combination. This method is called 'pretesting' and it suffers from the various problems mentioned above. One may also employ a factor-augmented regression model, but this approach is subject to the pretest for the number of factors, and the explanation of a concept is less clear if more than one factor is used. Extracting exactly one factor from each group can resolve the above problems, but it only makes use of a very limited part of information contained in the data. Since the procedure of extracting factors is independent of estimation procedure, it is not necessary that the extracted information is relevant and complete in explaining a specific dependent variable. On the contrary, our approach uses the full information of the data, and estimate the model in an integrate procedure (once the grouping is specified). Thus, we avoid the problems of factor analysis discussed above. More advanced methods such as Durlauf et al. (2008); Chipman (1996); Doppelhofer and Weeks (2009) are related with our approaches, but with significant difference which we shall discuss in details in Chapter 2.

We propose a hierarchical weighted least squares (HWALS) to address the measurement problem. In hierarchical model averaging, we introduce prior probabilities for the variables in each group, and treat the regression parameters as hierarchical random variables. We are uncertain about the error term, about which groups to select, and about which variables to select. All three levels of uncertainty are explicitly taken into account in our hierarchical model averaging estimation. Compared with existing methods, has several advantages. It provides an estimate and standard deviation for each group, which facilitates statistical inference and enables us to analyze the effect of each group; it combines model selection and estimation and thus avoids the problems associated with pretesting (see Danilov and Magnus (2004b) for a discussion and review of these problems); it allows researchers to assign various types of priors depending on the strength of their information and beliefs; it does not suffer from multicollinearity and dimensionality problems because it only considers models with one variable in each group; and its computational burden is very light, especially compared to standard Bayesian model averaging (BMA) and Bayesian averaging of classical estimates (BACE).

In Chapter 3 we study the prediction under uncertainty. We propose a weighted average least square (WALS) predictor, and and study its properties in simulations. The WALS procedure avoids some of the problems encountered in standard Bayesian model averaging (BMA). In particular, the prior is based on a coherent notion of ignorance, thus avoiding normality of the prior and unbounded risk. Also, the computational burden increases linearly rather than exponentially with the number of regressors, and is therefore trivial compared to other model averaging estimators such as standard BMA, model-selection-based weights methods (Buckland et al., 1997; Hjort and Claeskens, 2003), exponential reweighting (Yang, 2004), or Mallows model averaging (Hansen, 2007, 2008). Our proposed method explicitly allows for correlation in the observations, including possible correlation between the errors in the realized sample and the predictive sample.

Another main contribution of this chapter is that we propose an estimate for the prediction variance taking model uncertainty into account, and evaluate the accuracy of this estimate. Prediction variance is typically ignored in the existing model averaging prediction methods. In this chapter we evaluate the proposed prediction variance through simulation. We emphasize that the typical researcher's instinct favoring a predictor with a small variance over one with a large variance is not correct. We argue that what we require is not a small but a 'correct' variance: in a situation with much noise a predictor with a small variance can cause much harm, while a truthfully reported large variance may lead to more prudent policy. In fact, one of the problems with the credibility of econometric predictions may be that our reported prediction variances are too small, and this is caused, at least in part, by the fact that model uncertainty is ignored. We shall see that WALS predictions may lead to higher variances, but that these variances are closer to the truth.

1.2. Econometrics meets political economics

Policy-makers design policies to improve the economy, social welfare, or for other purposes. In reality, however, it is often that policies engender undesirable effects. Some policy may have the outcome that largely deviates from its original intention, or even the opposite outcome; More often, policies can have unexpected effects that harm other aspects of economy and social welfare which is not accounted. Therefore, policy-makers are especially interested in the effect a policy on the economy or social welfare. Chapter 4 and 5 study two influential policies in China and U.S., respectively, and investigate the effect of the policies on economy and social welfare.

Chapter 2 considers the policy 'West China Development Drive', introduced by the Chinese central government in 2000 with the purpose of stimulating the economy of the Western regions. This policy emphasizes on intensifying natural resource exploitation, and several significant projects connected with natural resources in West-China have resulted from this initiative. For example, the West-East natural gas transmission project led to an increase of natural gas production in Sichuan and Qinghai provinces by more than 100% and 900%, respectively, between 2000 and 2007. Also, steel production in Yunan and Guizhou provinces increased by around 200% and 400%, respectively, since the Drive began. The economic growth rate in Western provinces has indeed increased since 2000. We focus on how this policy affects the resource effect on economic growth. We first examine the interaction between natural resource, institutional quality, and economic growth using cross-sectional data. We propose several new measurements of resource abundance. These new measurements consider resource abundance either as a stock or as a flow, thus allowing a comparison between *in situ* resource reserves (a stock) and resource revenues (a flow, usually referred to as a 'windfall gain'), and they are regarded as less endogenous than conventional dependence measures. To allow the possible nonlinear interaction effect of institutional quality, we employ a functional coefficient model and find that the effect of resource abundance in China depends on institutional quality in a nonlinear fashion, which can not be fully captured using a linear model. More importantly, we find — in contrast to Mehlum et al. (2006) — that the effect of natural resources is more positive for provinces with poor institutional quality.

We then investigate whether resource effect is different before and after the policy by estimating a standard and a time-varying panel data model. Immediately after the 2000 policy shock, the positive correlation between economic growth and resource revenues was increasing, but this did not last long. After 2004 economic growth slowed down while resource revenues kept increasing, leading to weak correlation. The phenomena before 2004 is due to the emphasis of the West China Development Drive on exploiting the resources in the Western provinces more intensively and efficiently. Income in these regions has increased, stimulating economic growth, but not equally in all regions. The decreasing correlation between economic growth and resource revenues after 2004 shows that increased resource exploitation did not promote the development of other industries and sectors typically regarded as engines of economic growth. As exemplified in Figure 4.5, average R&D, industrialization, private sector employment, and foreign investment all changed relatively little as resource revenues increased sharply. Typical examples are Ningxia and Gansu provinces, where resource revenues increased significantly after 2000, but most of the other sectors were still underdeveloped. The economies of the Western provinces still relied much on primitive sectors, and the industrial structure of the Western provinces failed to modernize. The emphasis on resource exploitation brought extra income in the short run, but it did not narrow the gap between West and East China. In addition, only part of the resources produced by the Western regions was used to improve the local economy. The larger part was transported to the Eastern regions to meet the large demand for energy and resources there. For example, the most important gas field in Sichuan province transmitted more than 70% of its natural gas to Eastern provinces. This may also have resulted in enlarging the gap between Eastern and Western provinces.

Chapter 5 considers a fiscal policy, fiscal decentralization. Fiscal decentralization shifts some responsibilities for expenditures and/or revenues to lower levels government. During the last three decades, this policy has been at the center stage of policy experiments, not only in countries with a traditional tendency of decentralizing like United States, but also in a large number of developing and transition economies, such as Africa, Asia, and Latin America. Fiscal decentralization is widely believed to be an effective tool for improving the performance of public expenditure, partly because it introduces more competition, and also because lower levels of government have better knowledge of a citizen's demand. However, in this chapter we point out that fiscal decentralization also has a dark side that is currently not recognized in the literature: it can arise dress-up contests between local governments in which they compete for a better image, and such contests may lead to a structure bias in the public expenditure and further a social welfare loss.

We first setup a game between two local politicians and model dress-up contests. This model involves a concept of visibility proposed by Rogoff (1990) and Mani and Mukand (2007). Public goods are 'invisible' if their outcomes cannot well reflect the politicians' capability, either because they are difficult to observe or because they are determined by other factors beyond government's control. We study how fiscal decentralization affects the distribution of public expenditure on visible and invisible public goods, and we find that this policy distorts public expenditure by allocating more resources on the visible public goods.

Next, we empirically test the hypothesis implied by our theoretical model using the U.S. state level panel data. We focus on the poverty rate in U.S., a typical measure of social welfare, and treat cash assistance as a more visible public project and vendor payment as a less visible one. We find that fiscal decentralization causes a public expenditure flow from vendor payment to cash assistance, that is from a less visible project to a more visible one. This result provides evidence of the visibility effect, and also confirms our theoretical findings that fiscal decentralization can cause dress-up contests. To capture how the distortion of public expenditure affects poverty, we use a functional coefficient approach, and estimate a pooled panel and a panel with a fixed effect. This approach allows us to capture the possible nonlinear interaction between the ratio of cash to vendor payment, welfare expenditure, and poverty. We find that the distortion (measured by the ratio of cash to vendor payment) can largely weaken the impact of welfare expenditure on poverty reduction, and such influence appears to be nonlinear.

Chapter 2

Concept-based Bayesian model averaging and growth empirics[1]

2.1. Introduction

In applied econometrics, when estimating a regression equation, one has to decide which *concepts* (say inflation) to include in the regression: the 'specification' problem. In addition, one has to decide which *measurements* of these concepts to use (for example, CPI-based or PPI-based inflation): the 'measurement' problem. The measurement problem is common in practice because most economic variables can be measured in various ways. Climate, for example, as a potential determinant of growth, can be measured by the fraction of tropical zone, the tropical zone area, or the absolute latitude. Another example is the concept of market concentration, typically thought of as a factor that affects the financial stability of individual firms, which can be measured by the Herfindahl-Hirschman index or by the market share of the four largest firms.

The measurement problem raises at least three issues. First, different choices of measurements produce different estimates for the same concept, leading to ambiguity in explanation and policy implications. Second, multiple measurements typically cause multicollinearity if they are included in one regression model, so that the estimates for individual measurements lack precision, and statistical inference on a concept based on these estimates can therefore be misleading. Third, including multiple measurements in one model can also cause a problem of dimensionality when the number of explanatory variables is close to or even exceeds the number of observations.

The current paper addresses the measurement problem by introducing hierarchical (two-level) model averaging, where we perform model averaging over concepts *and* measurements. From here on we shall denote concepts as *groups*, and measurements as

[1]This chapter is coauthored with Jan R. Magnus.

variables. We propose a method called hierarchical weighted least squares (HWALS), a generalization of weighted average least squares (WALS) developed in Magnus et al. (2010). In hierarchical model averaging we introduce prior probabilities for the variables in each group, and treat the regression parameters as hierarchical random variables. We are uncertain about the error term, about which groups to select, and about which variables to select. All three levels of uncertainty are explicitly taken into account in hierarchical WALS estimation.

The HWALS procedure has several advantages. It provides an estimate and standard deviation for each group, which facilitates statistical inference and enables us to analyze the effect of each group; it combines model selection and estimation and thus avoids the problems associated with pretesting (see Danilov and Magnus (2004b) for a discussion and review of these problems); it allows researchers to assign various types of priors depending on the strength of their information and beliefs; it does not suffer from multicollinearity and dimensionality problems because it only considers models with one variable in each group; and its computational burden is very light, especially compared to standard Bayesian model averaging (BMA) and Bayesian averaging of classical estimates (BACE).

In the empirical growth literature the three types of uncertainty are especially important, because there is little consensus in this literature on which regressors to include, and, even if there is agreement on a regressor (group), there is still disagreement on which measurement (variable) of that regressor to use. In addition, the number of variables in growth empirics is large and may even exceed the number of observations. For example, Durlauf et al. (2005) listed 145 candidate regressors, while the number of countries is typically less in cross-country growth studies. Our paper employs HWALS to re-investigate the effects of the various growth determinants.

We mainly compare our estimates with those of Sala-i-Martin et al. (2004) and with the WALS estimates of Magnus et al. (2010). Our hierarchical model averaging estimates produce more intuitive signs and they are more robust. This is the benefit we gain from not ignoring the measurement problem, so that correlated variables within one group are not all included in the regression. Our empirical results also provide several new insights. For example, we find — in contrast to the current literature — that education and government intervention are not robust, because some of the variables in these groups have poor explanatory power in the growth regressions.

The paper is organized as follows. A literature review is provided in Section 2.2. In Section 5.2 we present the hierarchical estimation strategy. Section 4.3 describes the data, grouping, and scaling. We apply our estimation strategy to the data in Section 2.5 and discuss the results. Next we address the potential problem that the number of variables is too large to apply the HWALS technique directly. In that case, approximations are required and these are discussed in Section 2.6. Section 4.6 concludes. In our supplementary document (Magnus and Wang, 2013) we present extensions and more detailed analyses.

2.2. A brief review of the literature

The measurement problem is not new — it was mentioned, *inter alia*, in Brock et al. (2003) in the context of growth empirics. A popular method to deal with it is 'extreme bounds analysis' (Leamer and Leonard, 1983; Leamer, 1985), but this method has the disadvantage (in contrast to HWALS) that it produces various estimates for each concept. Another conventional method, called 'pretesting', is to try many different concepts and select the most appealing combination. There are many problems with this procedure (Danilov and Magnus, 2004) caused by the fact that model selection and estimation are completely separated, so that uncertainty in the model selection is ignored when reporting properties of the estimates. In contrast to pretesting, HWALS combines model selection and estimation in one procedure.

One may also employ a factor-augmented regression model. Here we must decide on the number of factors (pretest problem), and when more than one factor is used the explanation of a concept becomes more difficult. A possible solution to both problems is to extract just one factor from each group, but then we would use only a small portion of the information in the data.

Since Raftery et al. (1997), Bayesian model averaging has developed as a popular tool in addressing model uncertainty, especially in the application to the empirical growth literature; see, *inter alia*, Fernandez et al. (2001), Sala-i-Martin et al. (2004), and Ciccone and Jarociński (2010). Standard Bayesian model averaging addresses model uncertainty (which concepts to include) in growth regressions, while our approach addresses both model uncertainty and measurement uncertainty simultaneously, in the same spirit as Salimans (2012) who studied functional-form uncertainty and model uncertainty simultaneously.

A recent study by Durlauf et al. (2008) investigated the robustness of growth theories using Bayesian model averaging with a dilution prior. This is related to what we do, although the growth theories and their empirical proxies studied in Durlauf et al. (2008) differ in an essential way from our 'concepts' and 'measurements'. Multiple empirical proxies capture different aspects of a growth theory, and each aspect itself is a concept. For example, Durlauf et al. (2008) considered two proxies for the geography theory, namely the fraction of tropical/subtropical land area and the fraction of land near navigable water. These two proxies indeed measure two effects of geography on growth: climate and physical accessibility (two concepts), and for each concept they only have one measurement. The two proxies of the geography theory are not alternative measurements for the same concept (correlation is only around 0.14), and this is where their paper differs from ours. In our case, standard BMA cannot be applied, primarily because we do not allow our model space to contain models with multiple measurements within one group. The use of a dilution prior (George, 2010) captures the dilution property resulting from multicollinear variables, but it does not address the fact that multiple measurements of a concept are included in one model, leading to misleading Bayesian model averaging estimates (due to misleading likelihoods and estimates obtained from models containing multicollinear variables in the same group). By shrinking the model space, our HWALS procedure addresses this problem and also reduces the computational burden. HWALS thus also differs from the hierarchical dilution prior used in Durlauf et al. (2012) who worked with the whole model space.

Related to our approach is the jointness statistic proposed in Doppelhofer and Weeks (2009), which measures the dependence between explanatory variables. The jointness measure is the posterior probability that two or more variables appear in the same model. Multiple measurements of a concept are correlated with each other and are likely, but not certain, to have strong negative jointness. Conversely, variables that have negative jointness do not necessarily measure the same concept. Like other Bayesian approaches, the jointness measure computed from the posterior probability is also affected by the multicolliniearity of variables in the same group.

Our work is in the same spirit as the hierarchical structure studied by Brock et al. (2003) and the heredity prior proposed by Chipman (1996). Brock et al. (2003) employed a tree structure to construct prior probabilities, while Chipman (1996) considered priors for group predictors and for competing predictors. Our hierarchical averaging method resembles these two approaches, especially since all three methods average over a subset of models. But our method differs from the two approaches in at least four aspects. First, unlike Brock et al. (2003) who assigned equal and independent weights to each growth theory in a tree structure, HWALS allows for inequality and dependence between the various theories. Second, compared with the heredity prior, the method of restricting the model space is much simpler in HWALS (groups and variables). Third, our procedure allows us to assign various types of priors to measurements (imprecise priors, data-dependent priors) depending on the strength of the researcher's beliefs. Finally, HWALS provides an explicit form of the first two unconditional moments.

2.3. Hierarchical weighted average least squares

2.3.1. Groups and variables

We write the linear regression model as

$$y = X^*\beta^* + \epsilon = X_1^*\beta_1^* + X_2^*\beta_2^* + \epsilon, \tag{2.1}$$

where we note two deviations from standard notation. First we write X^* and β^* rather than X and β, because the regressors are considered to be 'groups', for example education or inflation. These are groups (concepts) rather than precisely defined variables. There are many measures of education and of inflation that the researcher could use. These measurements of the same concept in one group are our 'variables'. Second, we distinguish between focus regressors (labeled 1) and auxiliary regressors (labeled 2). Focus regressors are in the model irrespective of any preliminary test or diagnostic. These include the variables of specific interest and the variables that economic knowledge dictates to be in the model. Auxiliary regressors, on the other hand, may or may not be in the model, depending on prior knowledge and diagnostics.

We write the columns of the (group) regressors as

$$X_1^* = \left(x_{1,1}^*, \ldots, x_{1,k_1}^*\right), \qquad X_2^* = \left(x_{2,1}^*, \ldots, x_{2,k_2}^*\right), \tag{2.2}$$

and the components of the (group) parameter vectors as

$$\beta_1^* = \begin{pmatrix} \beta_{1,1}^* \\ \beta_{1,2}^* \\ \vdots \\ \beta_{1,k_1}^* \end{pmatrix}, \qquad \beta_2^* = \begin{pmatrix} \beta_{2,1}^* \\ \beta_{2,2}^* \\ \vdots \\ \beta_{2,k_2}^* \end{pmatrix}. \tag{2.3}$$

The distinction between *groups* and *variables* is important. The l_1-th focus group x_{1,l_1}^* contains m_{1,l_1} variables, and the l_2-th auxiliary group x_{2,l_2}^* contains m_{2,l_2} variables. Groups may contain only one variable. While the variables themselves are considered deterministic, a group is random (if there are at least two variables in the group) because the choice between the variables or the weighting scheme depends on the data (and on priors).

We attach prior probabilities to the variables based on our confidence. Thus,

$$\Pr\left(x_{1,l_1}^* = x_{1,l_1}^i\right) = \pi_{1,l_1}^i, \qquad \Pr\left(x_{2,l_2}^* = x_{2,l_2}^j\right) = \pi_{2,l_2}^j, \tag{2.4}$$

where $i = 1, \ldots, m_{1,l_1}$ and $j = 1, \ldots, m_{2,l_2}$, under the constraints

$$\sum_{i=1}^{m_{1,l_1}} \pi_{1,l_1}^i = 1, \qquad \sum_{j=1}^{m_{2,l_2}} \pi_{2,l_2}^j = 1. \tag{2.5}$$

Given specific variables x_{1,l_1}^i and x_{2,l_2}^j in each group, we construct the design matrices

$$X_1^{(i)} = \left(x_{1,1}^{i_1}, \ldots, x_{1,k_1}^{i_{k_1}}\right), \qquad X_2^{(j)} = \left(x_{2,1}^{j_1}, \ldots, x_{2,k_2}^{j_{k_2}}\right), \tag{2.6}$$

and the parameter vectors

$$\beta_1^{(i)} = \begin{pmatrix} \beta_{1,1}^{i_1} \\ \beta_{1,2}^{i_2} \\ \vdots \\ \beta_{1,k_1}^{i_{k_1}} \end{pmatrix}, \qquad \beta_2^{(j)} = \begin{pmatrix} \beta_{2,1}^{j_1} \\ \beta_{2,2}^{j_2} \\ \vdots \\ \beta_{2,k_2}^{j_{k_2}} \end{pmatrix}, \tag{2.7}$$

where $(i) = (i_1, \ldots, i_{k_1})$ and $(j) = (j_1, \ldots, j_{k_2})$. The resulting model can then be written as

$$y = X_1^{(i)}\beta_1^{(i)} + X_2^{(j)}\beta_2^{(j)} + \epsilon, \tag{2.8}$$

where we emphasize again that each model includes precisely *one* variable from each group.

2.3.2. A three-step procedure

Under the assumption that the prior distributions on separate groups are independent, the prior probability attached to a specific choice of variables (i) and (j) is given by

$$\pi^{(i,j)} = \prod_{l_1=1}^{k_1} \pi_{1,l_1}^{i_{l_1}} \prod_{l_2=1}^{k_2} \pi_{2,l_2}^{j_{l_2}}. \tag{2.9}$$

The validity of the independence assumption embodied in (2.9) depends on how the groups are set up, and it is therefore important to investigate the sensitivity of the results to different groupings. We consider this issue in Section 2.5.5. This is the first step.

For given (i) and (j) we estimate (2.8) by Bayesian model averaging. In Bayesian model averaging the estimates are computed as weighted averages of the estimates obtained over all possible models, thus allowing for the fact that auxiliary regressors may or may not be in the model, depending on priors and diagnostics. A major advantage of Bayesian model averaging is that it treats model selection and estimation as *one* procedure. We shall use a method called WALS (weighted average least squares), but this is not essential in the development. The advantages of WALS are both conceptual and computational. The version of WALS employed here is described in Magnus et al. (2010), and the estimates are made scale-independent using the weighting scheme proposed in De Luca and Magnus (2011). Thus we obtain the posterior mean (the WALS estimates),

$$b^{(i,j)} = \begin{pmatrix} \hat{\beta}_1^{(i,j)} \\ \hat{\beta}_2^{(i,j)} \end{pmatrix}, \tag{2.10}$$

and the posterior variance matrix $V^{(i,j)}$. This is the second step.

These posterior moments are, of course, still conditional on the choice of variables, that is, on (i) and (j). In the third and final step we obtain the unconditional posterior moments b and V from

$$b = \sum_{(i,j)} \pi^{(i,j)} b^{(i,j)} \tag{2.11}$$

and

$$V = \sum_{(i,j)} \pi^{(i,j)} \left(V^{(i,j)} + b^{(i,j)} b^{(i,j)\prime} \right) - bb'. \tag{2.12}$$

The variance V in the posterior distribution thus fully represents the three sources of uncertainty associated with the hierarchical procedure: uncertainty represented by the error term given the specification of the model; uncertainty about which auxiliary groups to include; and uncertainty about which variables to include in each group. The estimator b is the hierarchical WALS (HWALS) estimator, and V is taken to be its variance.

The HWALS estimator b cannot be interpreted as the usual marginal effect, since it corresponds to a group (concept) rather than to a variable (measurement). Since all variables are normalized to the same scale (see Section 4.3), the estimated coefficient of the i-th variable in a group is the *normalized* marginal effect, taking into account possible inclusion of other auxiliary variables. Due to the normalization, such effects are comparable not only within concepts but also between concepts. The averaged estimator (over the variables) of a group coefficient can thus be interpreted as the average effect of the group.

2.3.3. Choice of π

The prior probabilities π should be specified, and the question is how. The specification of π should depend on the strength of the researcher's prior information and beliefs on the 'quality' of the variables. We distinguish between four cases.

In the first case, the researcher has no prior information at all. In each group the quality of one variable is 'independent' of the quality of another, and equally good, so we assign equal weights within each group, that is,

$$\pi^i_{1,l_1} = \frac{1}{m_{1,l_1}}, \qquad \pi^i_{2,l_2} = \frac{1}{m_{2,l_2}}.$$

This is our default. Discrete uniform priors (over models) in a Bayesian model averaging framework were recently criticized by George (2010), especially in the presence of highly correlated regressors. He suggested the use of dilution priors in order to prevent the probability of a set of 'similar' models increasing when more similar variables are introduced. While this is a good idea, our case is different, because our prior probabilities are assigned to variables rather than the models, and thus the probabilities are not diluted by highly correlated variables.

In the second case, the researcher has no prior information but hopes to update the prior using the observed data. We propose to use data-dependent priors. We write $X_1^* = (X_{11} : X_{12}^*)$, where X_{11} contains the focus regressors for which only one variable is available, and X_{12}^* contains the focus regressors for which at least two variables are available. For each group l in X_{12}^* we estimate

$$y = X_{11}\beta_{11} + \beta_{1,l}^i x_{1,l}^i + \epsilon \qquad (i = 1, \dots, m_{1,l}), \tag{2.13}$$

from which we calculate the likelihood $L(x_{1,l}^i) = \Pr\left(y, X | x_{1,l}^* = x_{1,l}^i\right)$. Then we update the prior $\pi_{1,l}^i$ by Bayes' rule:

$$\bar{\pi}_{1,l}^i = \Pr\left(x_{1,l}^* = x_{1,l}^i | y, X\right) = \frac{\pi_{1,l}^i L(x_{1,l}^i)}{\sum_{h=1}^{m_{1,l}} \pi_{1,l}^h L(x_{1,l}^h)}.$$

A larger weight is thus assigned to the variable with more explanatory power (larger likelihood). Equation (2.13) is misspecified, because we ignore X_{12}^* (except one variable $x_{1,l}^i$) and all auxiliary regressors in X_2^*. However, the effect of the misspecification on $\bar{\pi}$ is partially 'divided out' and thus expected to be small.

Two subcases are of interest. In case 2(a) (one-step updating) we update the priors for the auxiliary variables in the same way, based on the equation

$$y = X_{11}\beta_{11} + \beta_{2,l}^j x_{2,l}^j + \epsilon \qquad (j = 1, \dots, m_{2,l}). \tag{2.14}$$

In case 2(b) (two-step updating) we update the priors for the auxiliary variables based on the extended equation

$$y = X_1^{(i)}\beta_1^{(i)} + \beta_{2,l}^j x_{2,l}^j + \epsilon \qquad (j = 1, \dots, m_{2,l}), \tag{2.15}$$

where *all* focus groups are used, not only the groups with one variable (X_{11}), but also the groups with two or more variables. For the latter we select the variable with the highest posterior probability $\bar{\pi}_{1,l}^i$.

The third case occurs when we have unequal prior information about the variables, and the exact values of prior probabilities are also known.

In the fourth case we can rank the prior probabilities within one group without knowing their precise values. Here we use 'imprecise probability' as our prior, namely $[\pi^-, \pi^+]$. This generalization of precise (point-valued) probability satisfies all principles of probability theory (Walley and Fine, 1982; Weichselberger, 2000), and allows us to model the uncertainty of subjective prior probabilities. The resulting estimates b and V are then also interval-valued.

2.4. Data, grouping, and scaling

We reexamine growth determinants using the proposed hierarchical method of Section 5.2. There is a large literature on explaining cross-country growth differences, but this literature has not led to a consensus on which regressors to include and on which measurement of that regressor to use. These issues are well exposed in Brock et al. (2003). Growth empirics thus provides a typical and important example of a situation where two types of uncertainty exist: uncertainty about the relevance of a group and uncertainty about which variable to select within the group.

Our data are taken primarily from Sala-i-Martin et al. (2004). The dependent variable is the average growth rate of GDP per capita 1960–1996. The Sala-i-Martin et al. (2004) data set contains 88 countries and 67 variables (plus the constant term). To this list we have added seven variables from Sala-i-Martin (1997): six variables in education and one variable in government intervention. These are indicated with an asterisk (*) in Table 2.1. This makes a total of 74 variables (25 groups) plus the constant term. We use 72 (rather than 88) countries, the maximum possible number if we wish to obtain a 'balanced' data set with an equal number of observations for all regressors. Since we have more variables than observations we cannot estimate the whole set. Grouping will therefore be especially helpful here. The issue of having more variables than observations has recently received new attention in the literature; see Huang et al. (2010) and Jensen and Würtz (2012) for alternative approaches.

Table 2.1: Grouping of variables: type I groups

g	Group	v	Variable
(1)	Demographic characteristics	1	Fraction population over 65
		2	Fraction population under 15
(2)	Economic system	3	Capitalism
		4	Socialism
(3)	Education	5	Primary schooling (1960 enrollment rate)
		6*	Secondary schooling (1960 enrollment rate)
		7	Higher education (1960 enrollment rate)
		8	Public education spending share in GDP in 1960s
		9*	Primary school years
		10*	Secondary school years
		11*	Higher education years
		12*	Average years of schooling
		13*	Average years of schooling $\times$ log of GDP per capita
(4)	Government intervention	14	Public investment share
		15*	Public consumption share (excl. education and defense)
		16	Government consumption share in 1960s
		17	Government share of GDP in 1960s
		18	Nominal government GDP share in 1960s
(5)	Health	19	Life expectancy in 1960
		20	Malaria prevalence in 1960s
(6)	Inflation	21	Average inflation 1960–1990
		22	Square of inflation 1960–1990
(7)	Initial state	23	GDP per capita in 1960 (log)
		24	Size of economy (GDP in 1960)
(8)**	Democracy	25	Political rights
		26	Civil liberties
(9)	Scale effect	27	Land area
		28	Population in 1960
(10)	Trade policy indices	29	Outward orientation
		30	Years open
(11)	Tropics effect	31	Fraction of tropical area
		32	Tropical climate zone
		33	Absolute latitude
(12)	War	34	Fraction spent in war 1960–1990
		35	War participation 1960–1990

* Variable is not in the Sala-i-Martin et al. (2004) data set, but taken from Sala-i-Martin (1997).
** Group (8) is called 'Democracy' following Barro (1999).

Table 2.2: Grouping of variables: type II groups

g	Group	v	Variable
(13)	Ethnicity and language	36*	Ethnolinguistic fractionalization
		37	English-speaking population
		38	Fraction speaking foreign language
(14)	Religion	39	Fraction Confucian
		40	Fraction Muslim
		41	Fraction Buddhist
		42	Fraction Protestant
		43	Fraction Hindu
		44	Fraction Catholic
		45	Fraction Orthodox
		46*	Religious intensity
(15)	Trade statistics	47*	Openness measure 1965–1974
		48	Primary exports in 1970
(16)	Terms of trade	49	Terms of trade ranking
		50	Terms of trade growth in 1960s
(17)	Regional effect	51	East Asian dummy
		52	African dummy
		53	European dummy
		54	Latin-American dummy
		55	Colony dummy
		56	British colony
		57	Spanish colony
(18)	Natural resources	58	Hydrocarbon deposits in 1993
		59	Fraction GDP in mining
		60	Oil-producing country dummy
(19)	Population	61	Population density coastal in 1960s
		62	Interior density
		63	Fraction population in tropics
		64	Population density in 1960
		65	Population growth rate 1960–1990
		66	Fertility in 1960s
(20)	Geography (excl. tropics effect)	67	Fraction land area near navigable water
		68	Landlocked country dummy
		69	Air distance to big cities
(21)	Price distortion	70	Investment price
(22)	Real exchange rate	71	Real exchange rate distortions
(23)	Defense	72	Defense spending share
(24)	Political instability	73	Revolutions and coups
(25)	Independence	74	Timing of independence

* Representative variable of the group.

The regressors are listed and grouped in Tables 2.1 and 2.2. The 74 variables are organized in 25 groups. The grouping is based on Durlauf et al. (2005) with two deviations: we split the 'geography' group in two ('tropics effect' and 'geography excluding tropics effect'), and we also split the 'government' group in two ('government intervention' and 'defense'). The reason is that within the new groups 'tropics effect' and 'government intervention' the same concept is measured, while the remaining items are of a different nature.

We distinguish between two types of groups. A group of type I (Table 2.1) contains variables providing alternative measurements of one concept. For example, a country's democracy (the concept) can be measured in several ways, and we allow two measurements (political rights and civil liberties). An important growth determinant is education (the concept), which attempts to capture human capital accumulation. Since the output of human capital investment is difficult to measure, one typically resorts to input variables, such as the enrollment rate, school years, or the share of public education spending. These input variables serve as different (but typically highly correlated) measurements for the same concept. We want to use only *one* measurement, but we do not know which one. Our theory of Section 5.2 applies to this type, that is, to groups (1)–(12).

In contrast, a group of type II (Table 2.2) contains variables measuring different aspects of one concept. For example, the group 'regional effect' contains seven dummy variables, each indicating whether a country belongs to some particular (colonial) region. These variables all measure a regional effect, but a different aspect of it, and these aspects are not highly correlated or easily aggregated. Our hierarchical theory does not apply to groups (13)–(20), because parameter estimates associated with these variables have different meanings, and hence a weighted sum of these estimates makes little sense. In our hierarchical estimation procedure we can either include all variables of a type II group or select a representative. For groups (13)–(15) we select one representative; for groups (16)–(20) we include all variables. Groups (21)–(25) only contain one variable, and hence there is no difference between variable and group. In summary, we have 12 type I groups (35 variables) and 13 type II groups (39 variables).

Grouping of variables can be ambiguous. While the grouping in Tables 2.1 and 2.2 based on Durlauf et al. (2005) is plausible, there is no complete agreement in the growth literature on how to group the large number of growth proxies. For example, one may argue that the enrollment rates and attainment levels in the education group may have different effects on growth, because the former relate to the flow of education (Mankiw et al., 1992) whereas the latter refer to stocks. We address such problems in Section 2.5.5.

Before we apply the hierarchical WALS procedure, we scale all variables, that is, we scale (and center) each variable x by replacing it with $(x - \text{mean}(x))/\text{std}(x)$, so that the resulting transformed variable has zero mean and unit variance. In standard (non-hierarchical) WALS the centering has no effect (other than on the constant term), but the scaling does. The latter effect can be removed by scaling the matrix

$$Z^{(i,j)} = X_2^{(j)\prime} \left(I - X_1^{(i)} \left(X_1^{(i)\prime} X_1^{(i)} \right)^{-1} X_1^{(i)\prime} \right) X_2^{(j)},$$

such that all its diagonal elements equal one (De Luca and Magnus, 2011). In hierarchical WALS the preliminary scaling is important because it makes the magnitudes of the estimated parameters within one group comparable.

In addition to scaling the variables, we may also wish to change the sign of some variables, so that variables within one group are positively correlated. For example, in the group 'health' we change the definition of malaria prevalence to malaria non-prevalence, so that both variables in this group now measure the same thing rather than opposite things. The five variables that have been re-signed are: fraction population over 65, socialism dummy, malaria prevalence, civil liberties, and absolute latitude. The within-group correlations are presented in Magnus and Wang (2013, Table 1).

2.5. Growth empirics

There is not much consensus in the empirical growth literature on which growth determinants are salient and robust among a large set of growth theories. Most papers report insignificant coefficients for most determinants. One reason is that growth theories are open-ended Brock and Durlauf (2001), another that the same concept can be measured by (sometimes many) different empirical proxies. In this paper we concentrate on the second aspect. Different choices of measurement may result in very different estimates. If we include all or many measurements of the same concept in one regression, then the t-ratios will be misleading due to multicollinearity. Our theory allows us to treat the 74 (plus the constant) different measurements (variables) as elements of only 25 (plus the constant) concepts (groups).

We have to choose which groups are focus and which auxiliary, and we shall discuss two variants. In variant HWALS-F1 only the constant term is a focus group, while all other groups are auxiliary. This is the typical model averaging framework in which all explanatory variables are allowed to be either included or excluded, as in Sala-i-Martin et al. (2004). More information is used in the second variant, HWALS-F8, where eight groups (including the constant term) are treated as focus groups. These groups are therefore included in every model. The eight focus groups consist of four type I groups (education, health, initial state, tropics effect), two type II groups with a representative variable (ethnicity and language, religion), and two type II groups with only one variable (price distortion, constant term). The distinction between focus and auxiliary is made at the group level: if a group is considered to be focus (auxiliary), then each variable in that group is also focus (auxiliary). Since the estimates in these two variants are highly similar, we only report results of HWALS-F8.

In Tables 2.3 and 2.4 we present the results for HWALS-F8 using uniform priors and data-dependent priors, and compare them with WALS-F8. The sensitivity of the results to using other priors is studied in Magnus and Wang (2013), where we also present the results for HWALS-F1. We find that the effects of proximate determinants on economic growth is robust to the choice of prior probability, except for the group education. The indirect effect (effect on other groups) of a different choice of prior probability is small, but the direct effect (effect on the group itself) varies across groups. In general, the choice of priors is not a serious issue for the estimation of the standard deviations in our growth empirics.

The WALS-F8 estimates are based on the 67 variables in Sala-i-Martin et al. (2004), hence without the seven additional variables from Sala-i-Martin (1997). They differ from those in Magnus et al. (2010, Table 7), because of the scaling and the different number of observations. The WALS-F8 estimates correspond to variables; the HWALS-F8 estimates to groups.

We shall also compare our results with the BACE estimates of Sala-i-Martin et al. (2004). Since the posterior moments given by BACE are conditional on inclusion, their precision is misleading as pointed out in Magnus et al. (2010). Therefore, we compare with the *unconditional* BACE moments according to Equations (8) and (14) in Sala-i-Martin et al. (2004). The full set of unconditional BACE estimates is available in Magnus and Wang (2013).

Table 2.3: HWALS and WALS estimates: focus variables

Variable	WALS-F8	HWALS-F8	
		Uniform prior	Data-dep. prior
Education		−0.0013 (0.0046)	0.0051 (0.0034)
5 Primary schooling	0.0037 (0.0188)		
6 Secondary schooling			
7 Higher education	−0.0079 (0.0081)		
8 Public edu. spending	−0.0007 (0.0160)		
9 Primary school yrs			
10 Secondary school yrs			
11 Higher education yrs			
12 Ave. school yrs			
13 Ave. school yrs × logGDP			
Health		0.0073 (0.0058)	0.0062 (0.0059)
19 Life expectancy	0.0144 (0.0271)		
20 Malaria prevalence	−0.0045 (0.0094)		
Initial state		−0.0045 (0.0064)	−0.0084 (0.0057)
23 GDP in 1960 (log)	−0.0073 (0.0168)		
24 Size of economy	0.0006 (0.0186)		
Tropics effect		−0.0030 (0.0034)	−0.0041 (0.0034)
31 Frac. of tropical area	0.0015 (0.0207)		
32 Tropical climate zone	0.0013 (0.0047)		
33 Absolute latitude	0.0054 (0.0195)		
Ethnicity and language			
36 Ethnolinguistic frac.	−0.0019 (0.0087)	−0.0030 (0.0028)	−0.0022 (0.0026)
37 English-speaking pop.	0.0014 (0.0053)		
38 Frac. foreign language	0.0006 (0.0062)		
Religion			
39 Fraction Confucian	0.0009 (0.0058)		
40 Fraction Muslim	−0.0004 (0.0079)		
41 Fraction Buddhist	0.0010 (0.0132)		
42 Fraction Protestant	−0.0122 (0.0161)		
43 Fraction Hindu	0.0003 (0.0074)		
44 Fraction Catholic	−0.0130 (0.0226)		
45 Fraction Orthodox	−0.0014 (0.0029)		
46 Religious intensity	−0.0035 (0.0095)	−0.0015 (0.0019)	−0.0022 (0.0018)
Price distortion			
70 Investment price	−0.0047 (0.0076)	−0.0041 (0.0017)	−0.0046 (0.0015)

Table 2.4: HWALS and WALS estimates: auxiliary variables

Variable	WALS-F8	HWALS-F8	
		Uniform prior	Data-dep. prior
Demographic characteristics		0.0027 (0.0048)	0.0026 (0.0044)
1 Frac. pop. over 65	−0.0011 (0.0204)		
2 Frac. pop. under 15	−0.0003 (0.0324)		
Economic system		−0.0010 (0.0016)	−0.0007 (0.0015)
3 Capitalism	0.0018 (0.0056)		
4 Socialism	−0.0000 (0.0067)		
Government intervention		−0.0003 (0.0021)	0.0005 (0.0020)
14 Public investment share	0.0016 (0.0044)		
15 Public consumption share (excl. education and defense)			
16 Gov. consumption share	−0.0367 (0.1602)		
17 Gov. share of GDP	0.0362 (0.1489)		
18 Nominal gov. GDP share	0.0001 (0.0078)		
Inflation		0.0005 (0.0022)	0.0004 (0.0019)
21 Average inflation	0.0042 (0.0179)		
22 Square of inflation	−0.0064 (0.0200)		
Democracy		0.0025 (0.0027)	0.0015 (0.0024)
25 Political rights	0.0047 (0.0102)		
26 Civil liberties	0.0002 (0.0075)		
Scale effect		0.0028 (0.0028)	0.0018 (0.0024)
27 Land area	0.0063 (0.0157)		
28 Population	0.0005 (0.0086)		
Trade policy indices		0.0009 (0.0024)	0.0009 (0.0026)
29 Outward orientation	−0.0008 (0.0055)		
30 Years open	−0.0032 (0.0104)		
War		0.0001 (0.0016)	−0.0003 (0.0015)
34 Frac. spent in war	0.0004 (0.0067)		
35 War participation	0.0022 (0.0086)		
Trade statistics			
47 Openness measure	−0.0004 (0.0147)	0.0006 (0.0029)	−0.0003 (0.0025)
48 Primary exports	−0.0026 (0.0104)		
Terms of trade			
49 Terms of trade ranking	0.0028 (0.0084)	0.0004 (0.0027)	0.0002 (0.0024)
50 Terms of trade growth	0.0026 (0.0058)	0.0035 (0.0024)	0.0021 (0.0022)
Regional effect			
51 East Asian dummy	0.0087 (0.0108)	0.0058 (0.0028)	0.0046 (0.0025)
52 African dummy	0.0017 (0.0117)	−0.0031 (0.0036)	−0.0020 (0.0032)
53 European dummy	0.0198 (0.0247)	0.0015 (0.0045)	0.0009 (0.0040)
54 Latin-American dummy	0.0125 (0.0258)	−0.0014 (0.0046)	−0.0002 (0.0042)
55 Colony dummy	−0.0023 (0.0155)	−0.0038 (0.0035)	−0.0040 (0.0031)
56 British colony	−0.0003 (0.0071)	0.0028 (0.0027)	0.0022 (0.0026)
57 Spanish colony	−0.0015 (0.0152)	0.0012 (0.0033)	0.0007 (0.0031)

Table 2.4: Continued

Natural resources			
58 Hydrocarbon deposits	0.0015 (0.0053)	0.0001 (0.0019)	0.0005 (0.0017)
59 Frac. GDP in mining	−0.0016 (0.0072)	−0.0013 (0.0019)	−0.0012 (0.0017)
60 Oil country dummy	−0.0020 (0.0052)	−0.0018 (0.0023)	−0.0004 (0.0021)
Population			
61 Population density coastal	0.0019 (0.0172)	0.0007 (0.0030)	0.0026 (0.0025)
62 Interior density	−0.0025 (0.0070)	−0.0010 (0.0017)	−0.0008 (0.0015)
63 Fraction pop. in tropics	0.0003 (0.0092)	0.0015 (0.0032)	0.0009 (0.0028)
64 Population density	−0.0032 (0.0060)	−0.0015 (0.0021)	−0.0009 (0.0018)
65 Population growth rate	0.0073 (0.0232)	0.0014 (0.0054)	0.0003 (0.0047)
66 Fertility	0.0007 (0.0224)	−0.0030 (0.0063)	−0.0006 (0.0052)
Geography (excl. tropics effect)			
67 Frac. land area near water	0.0018 (0.0118)	0.0016 (0.0032)	0.0001 (0.0030)
68 Landlocked country dummy	0.0027 (0.0040)	0.0002 (0.0018)	−0.0003 (0.0016)
69 Air distance to big cities	0.0009 (0.0102)	0.0010 (0.0025)	−0.0001 (0.0023)
Real exchange rate			
71 Real exchange rate dist.	−0.0031 (0.0107)	−0.0021 (0.0020)	−0.0019 (0.0019)
Defense			
72 Defense spending share	−0.0145 (0.0599)	−0.0004 (0.0017)	−0.0007 (0.0016)
Political instability			
73 Revolutions and coups	0.0043 (0.0064)	−0.0006 (0.0018)	−0.0003 (0.0017)
Independence			
74 Timing of independence	−0.0001 (0.0110)	0.0008 (0.0025)	0.0010 (0.0023)

2.5.1. Sign comparisons

Let us first compare the signs of HWALS-F8 using uniform priors with those of WALS-F8 and BACE, where we recall that the latter estimates are based on variables while the former are based on groups. We shall say that an HWALS estimate is 'totally different' from the BACE/WALS estimate if the sign of a type I group is opposite to *all* of its variables, and 'partially different' if the sign of a type I group is opposite to *some* of its variables. For type II groups this distinction is not necessary.

Comparing HWALS to WALS we see that in five of the type I groups the estimates are partially different, and in three type I and nine type II groups they are totally different. Hence, quite different estimation results are produced by HWALS as compared to WALS. The signs produced by HWALS are generally more intuitive than those produced by WALS, except for education. For example, HWALS suggests that regions with higher fractions of tropical land have lower growth rates, while all variables of the tropics effect have a positive sign in WALS. HWALS finds that being more open has a positive impact on growth, while all variables in the trade policy indices have a negative sign in WALS; and HWALS finds that African and Latin-American countries generally grow slower and

British colonial countries grow faster, while WALS reports the opposite. The HWALS estimates reflect the fact that 45% of Latin-American countries and 86% of Sub-Saharan African countries achieve growth rates below or around the first quartile, while 52% of British colonial countries achieve above-average growth rates.

For the group education, HWALS produces a negative (but not significant) estimate. This seems counterintuitive. Upon closer inspection we see that the education group contains many variables which are not robust and have relatively large standard deviations. We have nine education variables, and they measure education in three ways: the enrollment rate at different school levels (variables 5–7); educational attainment at different school levels (variables 9–13); and public spending on education (variable 8). Only the primary schooling enrollment rate in 1960 and the secondary school years have robust positive effects, while the signs of the remaining variables vary with the model specification. This is in line with most empirical growth literature, although some care needs to be taken in explaining the strongly positive estimate of primary schooling in 1960 (Barro and Lee, 1993). The variation between different measurements and the insignificance of most measurements lead to an insignificant estimate of the group education. Therefore, the education effect on growth does not appear to be as robust as some studies suggest.

Comparing HWALS to BACE, we find three type I groups that are partially different (education, democracy, trade policy indices), and nine variables of type II groups that are totally different. The signs produced by HWALS are more reliable and also more intuitive. For example, BACE produces opposite effects of political rights and civil liberties, while HWALS finds a negative effect ($b = 0.0025$, $t = 0.9$) of democracy (recall that civil liberties is re-signed), supporting recent studies (Barro, 1996). The HWALS estimate is in line with the HWALS estimates of political rights ($b = 0.0028$, $t = 1$) and civil liberties ($b = 0.0022$, $t = 0.9$) if we include each variable separately. However, if both variables are included simultaneously, then we obtain much smaller estimates of both variables and large variances due to high correlation. Also in contrast to BACE, HWALS finds positive correlation between growth and the European dummy, and concludes that a larger fraction of GDP in mining leads to a lower growth rate, which is supported by most cross-country studies on the 'resource curse'. Finally, countries with more land area near navigable water have access to more convenient transportation and are typically more open, thus enhancing growth, as shown by HWALS but not by BACE.

2.5.2. Precision comparisons

Next we compare the t-ratios produced by HWALS-F8, WALS, and BACE (unconditional moments). The WALS and BACE t-ratios are largely similar. HWALS is generally more precise than WALS and BACE, especially for those groups/variables that are typically thought of as robust determinants.

For the focus groups, HWALS reports $t = 1.26$ for health, while the t-ratios of the two health variables (life expectancy and malaria prevalence) are 0.53 and -0.48 in WALS; and 0.45 and -0.53 in BACE. In the group 'tropics effect', the t-ratios of its three variables vary greatly in both WALS and BACE. Only the fraction of tropical area has a t-ratio slightly larger than 1 (in absolute value) in BACE, while the other two measurements all have $|t| < 0.30$. WALS even reports a counterintuitive positive effect. In contrast, HWALS combines three variables and gives a t-ratio of this group of approximately 1. The estimate of 'ethnolinguistic fractionalization' produced by HWALS has $|t| = 1.07$, while WALS and BACE show $|t| = 0.22$ and $|t| = 0.30$, respectively.

For the auxiliary groups, most estimates of type I groups produced by HWALS are more significant than WALS and BACE (for example, demographic characteristics, inflation, and the scale effect). The estimates of most type II groups produced by HWALS are more significant than WALS and BACE.

2.5.3. Explanatory power

Particularly relevant is the contribution of various growth theories in explaining differences in cross-country growth rates. Since all variables are converted to the same scale, the estimates capture the explanatory power of each theory.

We find that investment price ($b = -0.0041$, $|t| = 2.4$) and the East Asian dummy ($b = 0.0058$, $|t| = 2.1$) are the most robust variables and explain much of the cross-country variation. Less robust but stronger in explanatory power is health ($b = 0.0073$, $|t| = 1.3$). Even less robust but still strong in explanatory power are initial state ($b = -0.0045$, $|t| = 0.7$) and the colony dummy ($b = -0.0038$, $|t| = 1.1$). These results provide evidence in favor of the neoclassical growth determinants, and they are also largely consistent with the findings in the conditional convergence literature and other related studies (Fernandez et al., 2001; Sala-i-Martin et al., 2004; Durlauf et al., 2008).

The groups tropics effect, ethnicity and language, African dummy, and terms of trade have slightly less explanatory power. Here our results differ from those in Sala-i-Martin et al. (2004) based on posterior inclusion probabilities. Unlike most studies, economic growth is not found to be robustly related to education and government intervention.

2.5.4. Data-dependent priors

The last column in Tables 2.3 and 2.4 presents HWALS-F8 estimates using data-dependent priors with one-step updating. (The updated priors and results of two-step updating are presented in Magnus and Wang (2013).) By construction, the two updating methods give the same updated prior probabilities for the focus variables, but they differ in the computation of the updated prior probabilities for the auxiliary variables. The differences are small except for demographic characteristics and the scale effect. The estimates produced by the two updating procedures generally have similar magnitudes and the same signs (except terms of trade ranking, Latin-American dummy, and fraction of land area near water). The exceptions all have a very weak effect on growth. The robustness of the updated probabilities and the resulting estimates confirms that model specification only has a marginal effect in the updating procedure.

We compare the HWALS-F8 results after updating the priors with the equal probability default. There is a big difference between focus and auxiliary groups. In the focus groups (especially education), the effects are generally different and stronger when the priors are updated than in the equal probability case. The reason lies in the fact that all focus groups have a dominant variable, while most auxiliary groups have equally important variables. For example, the large variation in updated prior probabilities (ranging from 0.978 to 0.003) in the group education shows that some variables in this group are much more relevant for economic growth than others. The ordering is generally in line with findings in other studies, e.g. Sala-i-Martin et al. (2004) and Magnus et al. (2010). Similarly, in the government intervention group, the government consumption share is by far the most relevant variable. Generally, the most relevant variables also have the highest posterior inclusion probability (Sala-i-Martin et al., 2004), or are the most significant (Magnus et al., 2010) compared to other variables in the same group.

In the auxiliary groups (such as democracy), the estimates and standard deviations when updating the priors are mostly in line with those using equal probabilities. As discussed above, this is because the variables in most auxiliary groups are almost equally important, so that their updated prior probabilities are close. Examples of such groups are economic system, inflation, and war. The variables in these groups are highly correlated, and hence including all variables in one regression leads to very insignificant estimates for some or all of the variables. For example, both Sala-i-Martin et al. (2004) and Magnus et al. (2010) reported extremely weak correlation between the capitalism/socialism dummy and economic growth. However, when considering them as a group, the correlation with growth is much stronger. Thinking in terms of groups rather than in terms of variables thus provides new insights.

2.5.5. Effect of different groupings

Our empirical results are based on the grouping in Tables 2.1 and 2.2. These groupings can of course be questioned and we briefly discuss the effect of alternative groupings. A more complete discussion is presented in our supplementary document (Magnus and Wang, 2013).

Initial state. In the group 'initial state' we separate the two variables GDP per capita in 1960 and the initial size of the economy, motivated by neoclassical growth model where initial GDP per capita has a structural role and thus should always be included (Mankiw et al., 1992). We thus treat GDP per capita in 1960 as a focus variable and the initial size of the economy as auxiliary. Since the initial level of income is now always included, the estimated coefficients should be interpreted as the effects of determinants of the height of the steady-state growth path, rather than as the effects of long-run growth determinants. The new grouping leads to a much larger estimated coefficient ($b = -0.0098$) of the initial level of income and a smaller variance ($V = 0.0053$), making initial income an important determinant and providing strong evidence of convergence. Results of other focus groups and most auxiliary groups are not largely affected.

Education. Education is a difficult concept to measure and our grouping can be easily criticized. We discuss four alternative groupings:

- Separate public education spending from the education group;
- Assign public education spending to the government intervention group;

- Distinguish between education flows and stocks by separating enrollment rates, attainment levels, and public education spending in three groups; and

- Distinguish between lower and higher education level by separating primary and secondary education, higher education, and public education spending in three groups.

The results confirm the large variation of education variables as well as their distinct effects on growth. Various aspects of education (flows versus stocks, lower versus higher level) are weakly related to growth, with the exception of primary schooling.

Tropics effect. Separating latitude from tropic effect group hardly affects the results.

2.6. Approximations for large k

To compute the HWALS estimates we need many runs of the WALS algorithm. Each run of the WALS algorithm requires model averaging over $k_2 = 41$ (HWALS-F1) or $k_2 = 34$ (HWALS-F8) auxiliary variables. In the case of BMA this would take much computing time (of the order 2^{k_2}), but in WALS much less (of the order k_2). This is one (but not the only one) advantage of WALS over BMA. Even so, in our application of the HWALS procedure, we have to repeat this algorithm $2^9 \times 3 \times 5 \times 9 = 69120$ times. This would be impossible with BMA or BACE, but it is still feasible in WALS, and the estimates reported in Tables 2.3 and 2.4 are based on exact computations.

If the number of groups and variables increases further, then estimating all combinations becomes computationally too time-consuming, especially if we also want to perform simulations and sensitivity analyses. In such cases we have to resort to approximations. In this section we propose and compare several approximating algorithms. There are two aspects to the approximation: selecting the subset of WALS regressions to be performed and obtaining the corresponding WALS estimates; and assigning estimates to the non-sampled regressions based on the estimates of the sampled regressions. We shall discuss each aspect in turn.

2.6.1. Subset selection

Two types of subset selection are considered: non-probability sampling and probability sampling. The non-probability method chooses the combinations deterministically. We sample those combinations whose prior probabilities (weights) are larger than a predetermined critical value π^*, because these are the combinations composed of relatively 'important' variables in each group. We obtain WALS estimates for these combinations. The 'precision' of the approximation is controlled by

$$\alpha = \sum_{\pi^{(i,j)} > \pi^*} \pi^{(i,j)},$$

representing the sum of the prior probabilities of the exact estimates used in the approximated HWALS computation. We use two stopping rules. First, we reduce π^* until the precision α satisfies a required level α^*. Second, to bound computation time, we restrict the number of samples S by an upper bound S^*. Hence, we require $\alpha > \alpha^*$ and $S < S^*$.

In contrast, the probability method uses the prior probabilities as weights and draws randomly (without replacement) based on these weights. Each combination can now be selected, but combinations with a high weight will have a higher selection probability than combinations with a low weight. The only requirement is $S < S^*$.

2.6.2. Approximating the non-sampled estimates

We consider two methods to approximate the non-sampled estimates from the sampled ones, first using neighboring estimates, then using a normalization of the probability. The first method is based on 'neighboring' estimates. For a given combination C, its 'neighbors' consist of those combinations containing at least one group represented by a variable that is also present in C. The approximation averages the neighboring estimates. Neighboring estimates are good approximates because changing the measurement of a group has a much smaller impact on estimates of other groups (indirect effect) than it does on the group itself (direct effect).

In the second method we normalize the probability of the sampled combinations, so that the sum of these probabilities equals 1, that is,

$$\pi_*^{(i,j)} = \frac{\pi^{(i,j)}}{\sum_{(m,n)\in\mathcal{C}} \pi^{(m,n)}}, \qquad (i,j) \in \mathcal{C}, \tag{2.16}$$

where $\mathcal{C}$ is the set of sampled combinations. From Equation (2.16) we see that estimates of more important samples contribute more to the approximates. The second method thus uses not only closely related information (neighboring estimates), but also less related information (non-neighboring estimates). It is not a priori clear whether this is good or bad, and we shall investigate the issue below.

2.6.3. Comparison of the methods

We now have four methods for the approximation procedure, as follows:

	Approximating method	
Sampling method	Ave. neighbor	Norm. probability
Non-probability	Method 1	Method 2
Probability	Method 3	Method 4

We compare the four methods from two aspects: approximation accuracy and computation time. For approximation accuracy our criterion is the average absolute deviation from the estimates obtained from the whole sample.

Figure 2.1: Approximation accuracy: 4 methods

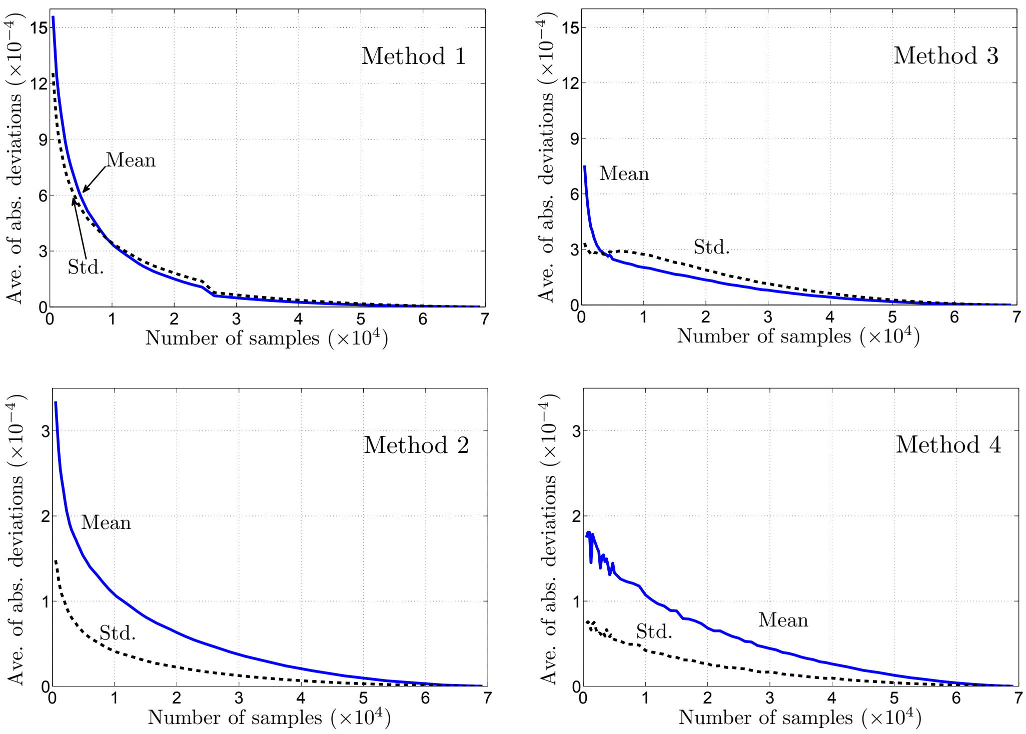


Figure 2.1 presents the convergence of the approximation accuracy for each of the four methods. Average absolute deviations decrease smoothly for non-probability methods, but less smoothly for probability methods because of the randomness. Comparing different approximating techniques, we find that Method 2 has higher approximation accuracy and needs less computation time than Method 1; and similarly that Method 4 has higher approximation accuracy and less computation time than Method 3. Apparently the normalization method strictly dominates the method using neighboring estimates, and this domination is especially strong when the number of samples is small. Next, when we compare different sampling techniques, we see that no method strictly dominates another. When the number of samples is small, Method 4 is more accurate than Method 2, but it is less accurate when the number of samples is large, thus reflecting the trade-off between using the more important estimates and a wider range of estimates.

The computation time is roughly proportional to the number of samples, so that computation time can be accurately predicted for each method. In fact, the ratio

$$\frac{\text{Computation time (in seconds)}}{\text{number of samples}/100}$$

is approximately 1.5 (Method 3), 1.2 (Method 4), 1.0 (Method 1), and 0.7 (Method 2). The computation time is higher for probability sampling than for non-probability sampling, because randomness is time-consuming. In summary, Methods 2 and 4 dominate Methods 1 and 3. When the number of samples is relatively small, Method 4 is preferred, but when the number of samples is relatively large, then Method 2 is preferred.

2.7. Conclusions

Applied researchers frequently encounter the situation where there is more than one measurement (variable) for a concept (group). To include all variables of the group into the regression is not satisfactory, because of multicollinearity. To choose between variables based on diagnostics leads to pretesting problems. A satisfactory solution can be obtained through two-level (hierarchical) Bayesian model averaging, where we question which groups should be in the model (level 1) and also which variables should be in each group (level 2). Our proposed method (HWALS) is an attempt to obtain estimates and standard deviations that fully reflect three sources of uncertainty: uncertainty represented by the error term, given the specification of the model; uncertainty about which (auxiliary) groups to include; and uncertainty about which variables to include in each group. Our method combines model selection and estimation and thus avoids the problem of pretesting. It is transparent, easy to implement, and computationally efficient compared to standard methods such as BMA and BACE. The method provides one estimate and standard deviation for each group (concept) rather than several estimates corresponding to each variable (measurement), and this facilitates statistical inference and interpretation of the effect of the concept. The hierarchical structure also allows us to assign various types of priors, depending on the strength of the researchers' beliefs. Unlike factor analysis, HWALS allows clear economic explanations, because the data are not transformed (except for simple scaling).

We apply the HWALS theory to growth empirics, and study the effects of different growth theories in explaining cross-country growth. This application is particularly suitable, because the open-ended growth theories and the many possible proxies for the same concept expose growth regressions to a high degree of model uncertainty. The HWALS estimates appear to possess more intuitive signs and are generally more significant compared to other methods. For example, HWALS finds a moderately negative effect of democracy and fraction GDP in mining, which is theoretically justified and empirically supported by the literature, whereas BACE finds the opposite result. Our findings regarding the robust and important determinants are mostly in line with the literature. A notable difference from the literature is that the education and government intervention effects are not robust, reflecting the large variation between variables in these two groups.

Extensive sensitivity analysis is provided with respect to the prior probabilities and grouping, from which we conclude that the main results, especially the estimates of robust and important determinants, are not sensitive. Also provided are methods of approximation when the number of groups or variables is large. The experimental results show that computation time can be much reduced while still obtaining estimates satisfying a given level of accuracy. Generalizations in various directions are possible. For example, non-linear models can be incorporated by adjusting the estimation method in the first-level averaging. The idea of hierarchical averaging can also be applied to other situations involving more than one level of uncertainty, such as model uncertainty with occasional structural breaks.

Chapter 3

Weighted average least square prediction[2]

3.1. Introduction

In econometric practice one typically first selects the 'best' model based on diagnostic tests (such as t-ratios, R^2, and various information criteria) and then computes estimates within this selected model. This is called 'pretesting' (Leeb and Pötscher, 2003, 2006, 2008). There are many problems with this procedure (Magnus, 1999; Danilov and Magnus, 2004b,a), but the most important is that model selection and estimation are completely separated so that uncertainty in the model selection is ignored when reporting properties of the estimates. Alternatively, one can consider averaging the results obtained from all candidate models, and this is called 'model averaging'. Model averaging not only appeals to common sense, but also has two major advantages. First, it avoids arbitrary thresholds (like 1.96), thus forcing continuity on a previously discontinuous estimator; second, it allows us to combine model selection and estimation into *one* procedure, thus moving from conditional to unconditional estimator characteristics.

Much of the model averaging literature has concentrated on estimation rather than on prediction. In this paper we concentrate on prediction (forecasting), which may in fact be a more appropriate application of model averaging, because the interpretation of coefficients changes with different models but the predictor always has the same interpretation. A substantial literature on the averaging of forecasts exists, going back to Bates and Granger (1969); see Granger (2003), Yang (2004), Elliott and Timmermann (2004), and Aiolfi and Timmermann (2006) for some recent contributions, and Hendry and Clements (2004) and Timmermann (2006) for recent reviews. Simulation and empirical studies indicate that predictors based on a set of models generally perform better than predictors obtained from a single model (Stock and Watson, 2004; Jackson and

[2]This chapter is coauthored with Jan R. Magnus and Xinyu Zhang.

Karlsson, 2004; Bjørnland et al., 2012).

Our paper has two main contributions. First, we introduce the prediction counterpart to the weighted average least squares (WALS) estimator proposed in Magnus et al. (2010) and study its properties in simulations. The WALS procedure avoids some of the problems encountered in standard Bayesian model averaging (BMA). In particular, the prior is based on a coherent notion of ignorance, thus avoiding normality of the prior and unbounded risk. Also, the computational burden increases linearly rather than exponentially with the number of regressors, and is therefore trivial compared to other model averaging estimators such as standard BMA, model-selection-based weights methods (Buckland et al., 1997; Hjort and Claeskens, 2003), exponential reweighting (Yang, 2004), or Mallows model averaging (Hansen, 2007, 2008). Our proposed method explicitly allows for correlation in the observations, including possible correlation between the errors in the realized sample and the predictive sample.

The second contribution of the paper is that we propose an estimate for the prediction variance taking model uncertainty into account, and evaluate the accuracy of this estimate. The typical researcher's instinct is to favor a predictor with a small variance over one with a large variance. We argue that what we require is not a small but a 'correct' variance: in a situation with much noise a predictor with a small variance can cause much harm, while a truthfully reported large variance may lead to more prudent policy. In fact, one of the problems with the credibility of econometric predictions may be that our reported prediction variances are too small, and this is caused, at least in part, by the fact that model uncertainty is ignored. We shall see that WALS predictions may lead to higher variances, but that these variances are closer to the truth.

The paper is organized as follows. Sections 3.2–3.7 develop the theory. In Section 3.2 we set up the model and present the traditional predictor. The commonly employed conditional predictor is presented in Section 3.3, and the WALS predictor in Section 3.4. In Section 3.5 we discuss the computation of the WALS predictor based on the Laplace prior. An estimator for the variance of the WALS predictor is proposed in Section 3.6. Finally, in Section 3.7, we discuss the estimation of unknown parameters in the variance matrix of the random disturbances. Then, in Sections 3.8–3.11, we compare the WALS predictor with its most important competitors: unrestricted maximum likelihood, pretesting, ridge regression, and Mallows model averaging. Our comparison is conducted through a large number of Monte Carlo simulation experiments, controlling for sample size, parameter values, and variance specifications. The simulation results show that the WALS predictor typically has the lowest mean squared prediction error among the predictors considered, and that the more uncertainty exists in the model, the better is the relative performance of WALS. Section 3.12 concludes.

3.2. The traditional predictor

Our framework is the linear regression model

$$y = X\beta + u, \tag{3.1}$$

where y is a vector of N observations on the dependent variable, X $(N \times k)$ is a matrix of regressors, u is a random vector of N unobservable disturbances, and β is a vector of k unknown parameters. We assume throughout that $1 \leq k \leq N-1$ and that X has full column-rank k. We are interested in some specific (possibly future) values of the regressors X_f $(N_f \times k)$, and we wish to predict the value y_f $(N_f \times 1)$ likely to be associated with X_f. We assume that y_f is generated by

$$y_f = X_f\beta + u_f, \tag{3.2}$$

and our task is to find a predictor $\hat{y}_f$ of y_f.

In general the observations will be correlated, and we shall assume that

$$\begin{pmatrix} u \\ u_f \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega & C_f' \\ C_f & \Omega_f \end{pmatrix} \right), \tag{3.3}$$

where the variance of (u, u_f) is a positive definite $(N+N_f) \times (N+N_f)$ matrix, whose component blocks Ω, C_f, and Ω_f are functions of an m-dimensional unknown parameter vector $\theta = (\theta_1, \dots, \theta_m)'$. Our theory applies to both fixed and random regressors under strictly exogeneity (hence not to lagged dependent variables). To simplify notation the following derivation treats the regressors as fixed (at least for the moment); the results for random regressors can be obtained similarly if we condition appropriately.

The joint distribution of u and u_f in (3.3) implies that

$$\mathrm{E}(u_f|u) = C_f\Omega^{-1}u, \qquad \mathrm{var}(u_f|u) = \Omega_f - C_f\Omega^{-1}C_f', \tag{3.4}$$

so that

$$\mathrm{E}(y_f|y) = X_f\beta + C_f\Omega^{-1}(y - X\beta). \tag{3.5}$$

This leads to the traditional least squares predictor in the presence of a non-scalar variance matrix:

$$\hat{y}_f = X_f\hat{\beta} + C_f\Omega^{-1}(y - X\hat{\beta}), \tag{3.6}$$

where $\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ denotes the generalized least squares (GLS) estimator of β, and it is assumed (for the moment) that θ is known; see Whittle (1963, p. 53, Eq. (10)) for the general formula, and Johnston and DiNardo (1997, Sec. 6.8) and Ruud (2000, Sec. 19.7) for the special case where $N_f = 1$ and the errors follow an AR(1) process. The predictor (3.6) is normally distributed with mean $\mathrm{E}(\hat{y}_f) = X_f\beta$ and variance

$$\mathrm{var}(\hat{y}_f) = X_f(X'\Omega^{-1}X)^{-1}X_f' + C_f(\Omega^{-1} - \Omega^{-1}X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1})C_f' \tag{3.7}$$

from which we see *inter alia* that the presence of the covariance C_f increases the variance of the predictor, and therefore that ignoring correlation leads to misleadingly precise predictions.

The prediction error $\mathrm{PE} := \hat{y}_f - y_f$ can be conveniently written as the sum of two independent random variables:

$$\mathrm{PE} = (X_f - C_f\Omega^{-1}X)(\hat{\beta} - \beta) - (u_f - C_f\Omega^{-1}u), \tag{3.8}$$

and the traditional predictor $\hat{y}_f$ is a good predictor in the sense that it is unbiased and that the prediction error has minimum variance

$$\begin{aligned}\mathrm{var}(\mathrm{PE}) = {} & (X_f - C_f\Omega^{-1}X)(X'\Omega^{-1}X)^{-1}(X_f - C_f\Omega^{-1}X)' \\ & + \Omega_f - C_f\Omega^{-1}C_f'\end{aligned} \tag{3.9}$$

in the class of linear unbiased estimators.

3.3. The conditional predictor

The previous section assumes that the data-generation process (DGP) and the model coincide, which one might call the 'traditional' approach. In practice, the model is likely to be (much) smaller than the DGP. In this section we shall assume that the model is a special case of the DGP obtained by setting some of the β-parameters equal to zero. We do not know in advance which β-parameters should be set to zero and we use model selection diagnostics (such as t- and F-statistics) to arrive at a model that we like. Once we have obtained this model we derive the properties of the predictor *conditional* on the selected model and hence we ignore the noise generated by the model selection process. We call this the 'conditional' approach. This is not quite right of course, and we shall present a method which combines model selection and prediction in the next section.

We distinguish between *focus* regressors X_1 (those we want in the model on theoretical or other grounds) and *auxiliary* regressors X_2 (those we are less certain of), and write model (3.1) accordingly as

$$y = X_1\beta_1 + X_2\beta_2 + u, \tag{3.10}$$

so that $X = (X_1 : X_2)$ and $\beta = (\beta_1', \beta_2')'$. Let $k_1 \geq 0$ be the dimension of β_1 and $k_2 \geq 0$ the dimension of β_2, so that $k = k_1 + k_2$. Model selection takes place over the auxiliary regressors only. Since each of the k_2 auxiliary regressors can either be included or not, we have 2^{k_2} models to consider.

In addition to the regressors that are always in the model (X_1) and those that are sometimes in the model (X_2), there are also regressors that are never in the model (say X_3), even though they are in the DGP. This is because the modeler is ignorant about these regressors or has no access to the necessary data. We disregard this situation for the moment, but return to it in Section 3.8.

We assume (at first) that θ and hence Ω is known. It is convenient to semi-orthogonalize the regression model as follows. Let

$$M_1^* := \Omega^{-1} - \Omega^{-1}X_1(X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}, \tag{3.11}$$

where we notice that the matrix $\Omega^{1/2}M_1^*\Omega^{1/2}$ is idempotent. Let P be an orthogonal matrix and Λ a diagonal matrix with positive diagonal elements such that $P'X_2'M_1^*X_2P = \Lambda$. Next define the transformed auxiliary regressors and the transformed auxiliary parameters as

$$X_2^* := X_2P\Lambda^{-1/2}, \qquad \beta_2^* := \Lambda^{1/2}P'\beta_2. \tag{3.12}$$

Then $X_2^*\beta_2^* = X_2\beta_2$, so that we can write (3.10) equivalently as

$$y = X_1\beta_1 + X_2^*\beta_2^* + u. \tag{3.13}$$

The result of this transformation is that the new design matrix $(X_1 : X_2^*)$ is 'semi-orthogonal' in the sense that $X_2^{*\prime} M_1^* X_2^* = I_{k_2}$ and this has important advantages that will become clear shortly.

3.3.1. Estimation in model $\mathcal{M}_i$

Our strategy will be to estimate (β_1, β_2^*) rather than (β_1, β_2). Each of the k_2 components of β_2^* can either be included or not included in the model and this gives rise to 2^{k_2} models. A specific model is identified through a $k_2 \times (k_2 - k_{2i})$ selection matrix S_i of full column-rank, where $0 \leq k_{2i} \leq k_2$, so that $S_i' = (I_{k_2 - k_{2i}} : 0)$ or a column-permutation thereof. Our first interest is in the GLS estimator of (β_1, β_2^*) in the i-th model, that is, in the GLS estimator of (β_1, β_2^*) under the restriction $S_i'\beta_2^* = 0$.

Let $\mathcal{M}_i$ represent model (3.13) under the restriction $S_i'\beta_2^* = 0$, and let $\hat{\beta}_{1(i)}$ and $\hat{\beta}_{2(i)}^*$ denote the GLS estimators of β_1 and β_2^* under $\mathcal{M}_i$. Extending Danilov and Magnus (2004a, Lemmas A1 and A2), the GLS estimators of β_1 and β_2^* under $\mathcal{M}_i$ may be written as (see also Magnus et al. et al, 2011):

$$\hat{\beta}_{1(i)} = (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}y - Q^*W_i b_2^*, \qquad \hat{\beta}_{2(i)}^* = W_i b_2^*, \tag{3.14}$$

respectively, where

$$b_2^* := X_2^{*\prime} M_1^* y, \quad Q^* := (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}X_2^*, \quad W_i := I_{k_2} - S_i S_i'. \tag{3.15}$$

Note that b_2^* is simply the GLS estimator of β_2^* in the unrestricted model, and that W_i is a diagonal $k_2 \times k_2$ matrix with k_{2i} ones and $(k_2 - k_{2i})$ zeros on the diagonal. The j-th diagonal element of W_i equals zero if β_{2j}^* (the j-th component of β_2^*) is restricted to zero, and equals one otherwise. If $k_{2i} = k_2$ then $W_i = I_{k_2}$. The diagonality of W_i is a direct consequence of the semi-orthogonal transformation.

The distributions of $\hat{\beta}_{1(i)}$ and $\hat{\beta}_{2(i)}^*$ are then

$$\hat{\beta}_{1(i)} \sim \mathrm{N}_{k_1}\left(\beta_1 + Q^* S_i S_i' \beta_2^*,\ (X_1'\Omega^{-1}X_1)^{-1} + Q^* W_i Q^{*\prime}\right), \tag{3.16}$$

$$\hat{\beta}_{2(i)}^* \sim \mathrm{N}_{k_2}\left(W_i \beta_2^*,\ W_i\right), \tag{3.17}$$

and $\mathrm{cov}(\hat{\beta}_{1(i)}, \hat{\beta}_{2(i)}^*) = -Q^* W_i$. The residual vector $e_i := y - X_1\hat{\beta}_{1(i)} - X_2^*\hat{\beta}_{2(i)}^*$ is given by $e_i = \Omega D_i^* y$, where $D_i^* := M_1^* - M_1^* X_2^* W_i X_2^{*\prime} M_1^*$ and $\Omega^{1/2} D_i^* \Omega^{1/2}$ is a symmetric idempotent matrix of rank $n - k_1 - k_{2i}$. It follows that:

- all models that include the j-th column of X_2^* as a regressor have the same estimator of β_{2j}^*, namely b_{2j}^*, irrespective of which other columns of X_2^* are included;
- the estimators $b_{21}^*, b_{22}^*, \dots, b_{2k_2}^*$ are independent; and
- the residuals of the i-th model $\mathcal{M}_i$ depend on y only through $M_1^* y$.

3.3.2. Prediction in model $\mathcal{M}_i$

Next we wish to predict N_f (possibly future) values y_f, based on values of the regressors X_{1f} ($N_f \times k_1$) and X_{2f} ($N_f \times k_2$). Corresponding to X_2^* we define $X_{2f}^* := X_{2f} P \Lambda^{-1/2}$, so that

$$\begin{pmatrix} y \\ y_f \end{pmatrix} = \begin{pmatrix} X_1 & X_2^* \\ X_{1f} & X_{2f}^* \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2^* \end{pmatrix} + \begin{pmatrix} u \\ u_f \end{pmatrix}, \tag{3.18}$$

where the errors (u, u_f) are distributed as in (3.3). From (3.5) we obtain

$$\mathrm{E}(y_f|y) = X_{1f}\beta_1 + X_{2f}^*\beta_2^* + C_f\Omega^{-1}(y - X_1\beta_1 - X_2^*\beta_2^*), \tag{3.19}$$

leading to the predictor in model $\mathcal{M}_i$, using (3.14),

$$\begin{aligned} \hat{y}_f^{(i)} &= X_{1f}\hat{\beta}_{1(i)} + X_{2f}^*\hat{\beta}_{2(i)}^* + C_f\Omega^{-1}(y - X_1\hat{\beta}_{1(i)} - X_2^*\hat{\beta}_{2(i)}^*) \\ &= X_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}y + C_f M_1^* y + Z_f W_i b_2^*, \end{aligned} \tag{3.20}$$

where

$$Z_f := (X_{2f}^* - X_{1f}Q^*) - C_f\Omega^{-1}(X_2^* - X_1Q^*). \tag{3.21}$$

The prediction error $\mathrm{PE}^{(i)} := \hat{y}_f^{(i)} - y_f$ can now be written as

$$\mathrm{PE}^{(i)} = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}u + Z_f(W_i b_2^* - \beta_2^*) - v_f, \tag{3.22}$$

where

$$Z_{1f} := X_{1f} - C_f\Omega^{-1}X_1, \qquad v_f := u_f - C_f\Omega^{-1}u. \tag{3.23}$$

Since v_f and u are uncorrelated, and $X_1'\Omega^{-1}u$ and b_2^* are also uncorrelated, we find that $\mathrm{PE}^{(i)}$ is the sum of three *independent* random variables.

Theorem 1: The prediction error $\mathrm{PE}^{(i)}$ follows a normal distribution with

$$\mathrm{E}(\mathrm{PE}^{(i)}) = -Z_f(I - W_i)\beta_2^*$$

and

$$\text{var}(\text{PE}^{(i)}) = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}Z_{1f}' + Z_f W_i Z_f' + \Omega_f - C_f\Omega^{-1}C_f',$$

and hence the mean squared prediction error $\text{MSPE}^{(i)} := \text{MSE}(\text{PE}^{(i)})$ is

$$\text{MSPE}^{(i)} = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}Z_{1f}' + Z_f \Delta_i Z_f' + \Omega_f - C_f\Omega^{-1}C_f',$$

where

$$\Delta_i := W_i + (I - W_i)\beta_2^*\beta_2^{*\prime}(I - W_i).$$

Proof: The results follow directly from (3.22). ‖

The best model is therefore the one where the matrix Δ_i is as 'small' as possible. Since W_i is a diagonal matrix with only zeros and ones on the diagonal, Δ_i is 'small' if the selected model $\mathcal{M}_i$ includes precisely those regressors x_{2j}^* of X_2^* whose corresponding parameter β_{2j}^* is larger than one in absolute value. Since the β_{2j}^* are 'theoretical' t-ratios, this result corresponds exactly to econometric intuition.

This econometric intuition is based on the following fact. Consider the partitioned regression model (3.10) with $\text{var}(u) = \sigma^2 I_n$. Let e_r denote the residual vector when y is regressed on X_1 only, and let e_u denote the residual vector when y is regressed on $X = (X_1 : X_2)$. Then, under the null hypothesis that $\beta_2 = 0$, the test statistic

$$F = \frac{(e_r'e_r - e_u'e_u)/k_2}{e_u'e_u/(N-k)} \tag{3.24}$$

is distributed as $F(k_2, N-k)$. This is a standard result (Johnston and DiNardo, 1997, p. 97). Now define the least squares estimator for σ^2 in the full model as $s_u^2 = e_u'e_u/(N-k)$, and the adjusted R^2 as $\bar{R}_u^2 = 1 - s_u^2/\sigma_y^2$, where $\sigma_y^2 := \sum_{n=1}^N (y_n - \bar{y})^2/(N-1)$. In the restricted model (where $\beta_2 = 0$) define s_r^2 and $\bar{R}_r^2$ accordingly. Then,

$$\frac{1-\bar{R}_r^2}{1-\bar{R}_u^2} = \frac{s_r^2}{s_u^2} = 1 + \frac{k_2}{N-k_1}(F-1), \tag{3.25}$$

and hence

$$s_u^2 < s_r^2 \iff \bar{R}_u^2 > \bar{R}_r^2 \iff F > 1. \tag{3.26}$$

As a special case ($k_2 = 1$), we find that adding one regressor will decrease s^2 and increase $\bar{R}^2$ if and only if the t-statistic of the corresponding parameter is larger than one in absolute value.

3.4. The WALS predictor

The problem, of course, is that we don't know which model to choose. Given estimates $\hat{\beta}^*_{2j}$ of the k_2 components β^*_{2j} of β^*_2, we could include the regressor x^*_{2j} if $|\hat{\beta}^*_{2j}| > 1$, and exclude it otherwise. This would lead to a *pretest* estimator with well-established poor properties. These poor properties stem primarily from the fact that the pretest estimator is 'kinked'; it has a discontinuity at one. This is not only mathematically undesirable but also intuitively: If $\hat{\beta}^*_{2j} = 0.99$ we exclude x^*_{2j}; if $\hat{\beta}^*_{2j} = 1.01$ we include it. It would seem better to include x^*_{2j} 'continuously' in such a way that the higher is $|\hat{\beta}^*_{2j}|$, the more of x^*_{2j} is included in our model. This is precisely the idea behind model averaging. The additional benefit of model averaging is that we develop the theory taking into account both model uncertainty and parameter uncertainty. In other words, we think of model selection and parameter estimation as *one* combined procedure, so that the reported standard errors reflect both types of uncertainty.

Thus motivated, we define the WALS predictor of y_f as

$$\hat{y}_f = \sum_{i=1}^{2^{k_2}} \lambda_i \hat{y}_f^{(i)}, \tag{3.27}$$

where the sum is taken over all 2^{k_2} different models obtained by setting a subset of the β^*_2's equal to zero, and the λ_i's are weight-functions satisfying certain minimal regularity conditions, namely

$$\lambda_i \geq 0, \quad \sum_{i=1}^{2^{k_2}} \lambda_i = 1, \quad \lambda_i = \lambda_i(M^*_1 y). \tag{3.28}$$

The assumption that the weights λ_i depend only on $M^*_1 y$ is justified by the fact that these weights are typically chosen according to the diagnostic power of the auxiliary regressors. This suggests that we choose $\lambda_i = \lambda_i(\hat{\beta}^*_{2(i)}, s^2_u)$, similar to the choice of weights in Liang et al. (2011). Since $\hat{\beta}^*_{2(i)}$ and s^2_u are both functions of $M^*_1 y$, it follows that the λ_i depend only on $M^*_1 y$. This condition significantly alleviates the computational burden. The definition (3.27) then specializes as follows.

Definition 1 (WALS predictor): The WALS predictor of y_f is given by

$$\hat{y}_f := X_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}y + C_f M^*_1 y + Z_f \hat{\beta}^*_2,$$

where $\hat{\beta}^*_2 := W b^*_2$ and $W := \sum_i \lambda_i W_i$.

Note that, while the W_i's are non-random diagonal matrices, the matrix W is random (but still diagonal) because it depends on the random λ_i's. The prediction error PE := $\hat{y}_f - y_f$ now takes the form

$$\mathrm{PE} = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}u + Z_f(\hat{\beta}_2^* - \beta_2^*) - v_f, \tag{3.29}$$

and we present its moments in the following 'equivalence' theorem.

Theorem 2 (Equivalence theorem): If the weights λ_i satisfy condition (3.28), then the WALS prediction error PE has the following expectation, variance and mean squared error:

$$\mathrm{E(PE)} = Z_f\,\mathrm{E}(\hat{\beta}_2^* - \beta_2^*),$$

$$\mathrm{var(PE)} = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}Z_{1f}' + Z_f\,\mathrm{var}(\hat{\beta}_2^*)Z_f' + \Omega_f - C_f\Omega^{-1}C_f',$$

and hence

$$\mathrm{MSE(PE)} = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}Z_{1f}' + Z_f\,\mathrm{MSE}(\hat{\beta}_2^*)Z_f' + \Omega_f - C_f\Omega^{-1}C_f'.$$

Proof: The key ingredient is that $\mathrm{cov}(M_1^*u, X_1'\Omega^{-1}u)$ and $\mathrm{cov}(u, v_f)$ are both zero. In addition, the λ_i (and hence W) depend only on M_1^*y so that $\hat{\beta}_2^* = Wb_2^*$ also depends only on M_1^*y. Hence, the three random variables $X_1'\Omega^{-1}u$, $\hat{\beta}_2^*$, and v_f are all independent of each other. The results follow. ‖

The equivalence theorem tells us that the WALS predictor $\hat{y}_f$ will be a 'good' predictor of y_f in the mean squared error sense if and only if $\hat{\beta}_2^*$ is a 'good' estimator of β_2^*. That is, if we can find λ_i's such that $\hat{\beta}_2^*$ is an 'optimal' estimator of β_2^*, then *the same* λ_i's will provide an 'optimal' predictor of y_f.

Next we obtain expressions for the bias and variance of the predictor itself, under the assumption that the diagonal elements of W depend only on $b_2^* = X_2^{*\prime}M_1^*y$ rather than only on M_1^*y.

Theorem 3: If the diagonal elements w_j of W depend only on b_2^*, then the WALS predictor $\hat{y}_f$ has the following bias and variance:

$$\mathrm{E}(\hat{y}_f - X_{1f}\beta_1 - X_{2f}\beta_2) = Z_f\,\mathrm{E}(\hat{\beta}_2^* - \beta_2^*)$$

and

$$\operatorname{var}(\hat{y}_f) = X_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_{1f}' + C_f M_1^* C_f' + Z_f \operatorname{var}(\hat{\beta}_2^*)Z_f'$$
$$+ C_f M_1^* X_2^* \operatorname{cov}(b_2^*, \hat{\beta}_2^*)Z_f' + Z_f \operatorname{cov}(\hat{\beta}_2^*, b_2^*)X_2^{*\prime} M_1^* C_f'.$$

Under the stronger assumption that w_j depends only on b_{2j}^*, the $k_2 \times k_2$ matrices $\operatorname{var}(\hat{\beta}_2^*)$ and $\operatorname{cov}(b_2^*, \hat{\beta}_2^*)$ are both diagonal.

Proof: The bias follows directly from Theorem 2. Noting that

$$\operatorname{cov}(X_1'\Omega^{-1}y,\ M_1^*y) = X_1'M_1^* = 0, \qquad \operatorname{cov}(X_1'\Omega^{-1}y,\ \hat{\beta}_2^*) = 0,$$

Definition 1 implies that

$$\operatorname{var}(\hat{y}_f) = X_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_{1f}' + C_f M_1^* C_f' + Z_f \operatorname{var}(\hat{\beta}_2^*)Z_f'$$
$$+ C_f \operatorname{cov}(M_1^*y, \hat{\beta}_2^*)Z_f' + Z_f \operatorname{cov}(\hat{\beta}_2^*, M_1^*y)C_f'.$$

Since $\Omega^{1/2}M_1^*\Omega^{1/2}$ is idempotent, we can write

$$\Omega^{1/2}M_1^*\Omega^{1/2} = AA', \qquad A'A = I_{n-k_1}.$$

Define $y^* := A'\Omega^{-1/2}y$ and $B_1 := A'\Omega^{-1/2}X_2^*$, so that $y^* \sim \mathrm{N}(B_1\beta_2^*, I_{n-k_1})$. Since $B_1'B_1 = I_{k_2}$ there exists an $(n-k_1) \times (n-k)$ matrix B_2, such that $B := (B_1 : B_2)$ is orthogonal. This allows us to write

$$M_1^*y = \Omega^{-1/2}A(B_1B_1' + B_2B_2')y^*, \qquad \hat{\beta}_2^* = WB_1'y^*,$$

so that

$$\begin{aligned}
\operatorname{cov}(M_1^*y, \hat{\beta}_2^*) &= \operatorname{cov}(\Omega^{-1/2}AB_1B_1'y^*, WB_1'y^*) + \operatorname{cov}(\Omega^{-1/2}AB_2B_2'y^*, WB_1'y^*) \\
&= M_1^*X_2^* \operatorname{cov}(b_2^*, \hat{\beta}_2^*) + \Omega^{-1/2}AB_2 \operatorname{cov}(B_2'y^*, WB_1'y^*) \\
&= M_1^*X_2^* \operatorname{cov}(b_2^*, \hat{\beta}_2^*),
\end{aligned}$$

because $B_1'y^*$ and $B_2'y^*$ are independent, and the diagonal elements w_j of W depend only on $X_2^{*\prime}M_1^*y = B_1'y^*$.

Finally, if w_j depends only on b_{2j}^*, then

$$\operatorname{cov}(b_{2i}^*, w_jb_{2j}^*) = 0, \qquad \operatorname{cov}(w_ib_{2i}^*, w_jb_{2j}^*) = 0 \qquad (i \neq j),$$

because b_{2i}^* and b_{2j}^* are independent. In that case both $\operatorname{cov}(b_2^*, \hat{\beta}_2^*)$ and $\operatorname{cov}(\hat{\beta}_2^*, b_2^*)$ are diagonal. This completes the proof. $\|$

3.5. Computation of the WALS predictor based on prior ignorance

The WALS predictor proposed in Definition 1 can not be computed unless we know $W = \sum_i \lambda_i W_i$. Because of the semi-orthogonal transformation, we do know that W is diagonal, say $W = \text{diag}(w_1, \ldots, w_{k_2})$. There are 2^{k_2} λ_i's, but there are only k_2 w_j's. These are functions of the λ_i's, but we can not identify the λ_i's from the w_j's. This does not matter because we are not interested in the λ_i's as we are not interested in selecting the 'best' model. We are only interested in the 'best' predictor.

The k_2 components b^*_{2j} of b^*_2 are independent with $\text{var}(b^*_{2j}) = 1$. Therefore, if we choose w_j to be a function of b^*_{2j} only, then the components $\hat{\beta}^*_{2j} = w_j b^*_{2j}$ of $\hat{\beta}^*_2$ will also be independent, and our k_2-dimensional problem reduces to k_2 one-dimensional problems. The one-dimensional problem is simply how to estimate β^*_{2j} using only the information that $b^*_{2j} \sim \text{N}(\beta^*_{2j}, 1)$.

This seemingly trivial question was addressed in Magnus (2002), who proposed the 'Laplace' estimator. This estimator is obtained by combining the normal likelihood with the Laplace prior,

$$b^*_{2j} \mid \beta^*_{2j} \sim \text{N}(\beta^*_{2j}, 1), \qquad \pi(\beta^*_{2j}) = (c/2)\exp(-c|\beta^*_{2j}|), \tag{3.30}$$

where c is a positive constant. The Laplace estimator is now defined as the resulting posterior expectation $\hat{\beta}^*_{2j} := \text{E}(\beta^*_{2j}|b^*_{2j})$. It is admissible, has bounded risk, has good properties around $|\beta^*_{2j}| = 1$, and is near-optimal in terms of minimax regret. It is also easily computable. The mean and variance of $\beta^*_{2j}|b^*_{2j}$ are given in Theorem 1 of Magnus et al. (2010). The mean is

$$\hat{\beta}^*_{2j} = \text{E}(\beta^*_{2j} \mid b^*_{2j}) = b^*_{2j} - c \cdot h(b^*_{2j}) \tag{3.31}$$

with

$$h(x) := \frac{e^{-cx}\Phi(x-c) - e^{cx}\Phi(-x-c)}{e^{-cx}\Phi(x-c) + e^{cx}\Phi(-x-c)}, \tag{3.32}$$

and the variance $v_j := \text{var}(\beta^*_{2j}|b^*_{2j})$ is

$$v_j = v(b^*_{2j}) = 1 + c^2(1 - h^2(b^*_{2j})) - \frac{c(1 + h(b^*_{2j}))\phi(b^*_{2j} - c)}{\Phi(b^*_{2j} - c)}, \tag{3.33}$$

where ϕ and Φ denote the density function and the cumulative distribution function of the standard-normal distribution, respectively.

The weights w_j are defined implicitly by $\hat{\beta}^*_{2j} = w_j b^*_{2j}$ and are thus given by

$$w_j = w(b^*_{2j}) = 1 - \frac{c \cdot h(b^*_{2j})}{b^*_{2j}}. \tag{3.34}$$

Each w_j satisfies $w(-b^*_{2j}) = w(b^*_{2j})$ and increases monotonically between $w(0)$ and $w(\infty) = 1$. Hence, $\hat{\beta}^*_{2j}$ is a shrinkage estimator, and we have

$$w(0)|b^*_{2j}| < |\hat{\beta}^*_{2j}| < |b^*_{2j}|. \tag{3.35}$$

In particular, when $c = \log 2$, we find that $w(0) = 0.5896$ which defines the maximum allowable shrinkage.

The hyperparameter c is chosen as $c = \log 2$, because this implies

$$\Pr(\beta^*_{2j} > 0) = \Pr(\beta^*_{2j} < 0), \qquad \Pr(|\beta^*_{2j}| > 1) = \Pr(|\beta^*_{2j}| < 1). \tag{3.36}$$

What this means is that we assume a priori ignorance about whether β^*_{2j} is positive or negative, and also about whether $|\beta^*_{2j}|$ is larger or smaller than one. These seem natural properties for a prior in our context, because we don't know a priori whether the β^*_2 coefficients are positive or negative, and we don't know either whether adding a specific column of X^*_2 to the model will increase or decrease the mean squared error of the predictors. Such a prior thus captures prior ignorance in a natural way. Given the choice of the weights w_j and hence of the estimator $\hat{\beta}^*_2$, the WALS predictor $\hat{y}_f$ can be computed.

3.6. Moments of the WALS predictor

The moments of the WALS predictor are given in Theorem 3, but the expressions provided there depend on unknown quantities. Under the assumption that the weights w_j are specified as in (3.34), and hence depend on b^*_{2j} only, we estimate these unknown quantities as follows.

Theorem 4: If the diagonal elements w_j of W depend only on b^*_{2j} as specified in (3.34), then the expected bias of the WALS predictor $\hat{y}_f$, based on prior densities $\pi(\beta^*_{2j})$, is zero:

$$\mathrm{E}\left(\mathrm{E}(\hat{y}_f - X_{1f}\beta_1 - X_{2f}\beta_2)|\beta^*_2\right) = 0.$$

Proof: According to Theorem 3, the prediction bias, conditional on β_2^*, is

$$\mathrm{E}(\hat{y}_f - X_{1f}\beta_1 - X_{2f}^*\beta_2^*|\beta_2^*) = Z_f\,\mathrm{E}(\hat{\beta}_2^* - \beta_2^*|\beta_2^*).$$

Further,

$$\begin{aligned}\mathrm{E}(\hat{\beta}_{2j}^* - \beta_{2j}^*) &= \mathrm{E}\left(\mathrm{E}(\hat{\beta}_{2j}^* - \beta_{2j}^*|\beta_{2j}^*)\right)\\ &= \mathrm{E}\left(\mathrm{E}(b_{2j}^* - \beta_{2j}^*|\beta_{2j}^*)\right) - c\cdot\mathrm{E}\left(\mathrm{E}(h(b_{2j}^*)|\beta_{2j}^*)\right) = 0,\end{aligned}$$

because $\mathrm{E}(h(b_{2j}^*)|\beta_{2j}^*)$ is antisymmetric in β_{2j}^* and $\pi(\beta_{2j}^*)$ is symmetric in β_{2j}^*. Hence the expected bias of $\hat{y}_f$ vanishes. $\|$

The variance of $\hat{y}_f$ is given in Theorem 3. Under the assumption that the weights w_j depend only on b_{2j}^*, the matrices $\mathrm{var}(\hat{\beta}_2^*)$ and $\mathrm{cov}(b_2^*, \hat{\beta}_2^*)$ are both diagonal. Hence it suffices to discuss the estimation of $\mathrm{var}(\hat{\beta}_{2j}^*)$ and $\mathrm{cov}(b_{2j}^*, \hat{\beta}_{2j}^*)$. The variance in the posterior distribution of $\beta_{2j}^*|b_{2j}^*$ is given by v_j in (3.33), and hence provides the obvious estimate of $\mathrm{var}(\hat{\beta}_{2j}^*)$. It is less obvious how to find an appropriate estimate of $\mathrm{cov}(b_{2j}^*, \hat{\beta}_{2j}^*)$. We propose

$$w_j = \widehat{\mathrm{cov}}(b_{2j}^*, \hat{\beta}_{2j}^*) = \widehat{\mathrm{cov}}(b_{2j}^*, w(b_{2j}^*)b_{2j}^*). \tag{3.37}$$

Since $\mathrm{var}(b_{2j}^*) = 1$, this would be a perfect estimate if w_j were a constant. Now, w_j depends on b_{2j}^* and is therefore not a constant. Still, its variation is very small compared to the variation in b_{2j}^*. The correlation associated with the covariance is

$$\widehat{\mathrm{corr}}(b_{2j}^*, \hat{\beta}_{2j}^*) = \frac{\widehat{\mathrm{cov}}(b_{2j}^*, \hat{\beta}_{2j}^*)}{\sqrt{\widehat{\mathrm{var}}(b_{2j}^*)\widehat{\mathrm{var}}(\hat{\beta}_{2j}^*)}} = \frac{w(b_{2j}^*)}{\sqrt{v(b_{2j}^*)}}, \tag{3.38}$$

since we estimate $\mathrm{var}(\hat{\beta}_{2j}^*)$ by $v_j = v(b_{2j}^*)$. The estimated correlation is therefore always positive (in fact, larger than 0.7452) and smaller than one, such that when b_{2j}^* approaches $\pm\infty$ the correlation approaches one.

We conclude that a suitable estimator for the variance of the WALS predictor is given by

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{y}_f) = X_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_{1f}' + C_f M_1^* C_f' + Z_f V Z_f'\\ + C_f M_1^* X_2^* W Z_f' + Z_f W X_2^{*\prime} M_1^* C_f',\end{aligned} \tag{3.39}$$

where V and W are diagonal $k_2 \times k_2$ matrices whose j-th diagonal elements v_j and w_j are given in (3.33) and (3.34), respectively. Having thus obtained estimators for all unknown quantities, the prediction variance can be computed.

3.7. Unknown variance matrix

We have thus far assumed that Ω and C_f are known, whereas in practice they are of course unknown. If the structure of the variance matrix is known, we can estimate Ω and C_f once we have an estimate of unknown parameter θ. The parameter θ can be estimated based on the unrestricted model by minimizing

$$\varphi(\theta) := \log |\Omega| + y'(\Omega^{-1} - \Omega^{-1}X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1})y \tag{3.40}$$

with respect to θ.

This leads to the maximum likelihood estimator $\hat{\theta}$ of θ, and hence to the estimators $\hat{\Omega} = \Omega(\hat{\theta})$ and $\hat{C}_f = C_f(\hat{\theta})$. Note that the gradient of φ is the $m \times 1$ vector whose i-th component is given by

$$\frac{\partial \varphi(\theta)}{\partial \theta_i} = \operatorname{tr}\left(\Omega^{-1}\frac{\partial \Omega}{\partial \theta_i}\right) - (M^* y)'\frac{\partial \Omega}{\partial \theta_i}(M^* y), \tag{3.41}$$

where

$$M^* = M_1^*(\Omega - X_2^* X_2^{*\prime})M_1^*. \tag{3.42}$$

Therefore, $\hat{\theta}$ depends on y only through $M_1^* y$ and the same holds for $\hat{\Omega}$ and $\hat{C}_f$. Replacing the unknown parameters with their estimates can have an effect on the property of the WALS predictor. However, Danilov (2005) showed that such an effect is small, at least in terms of the risk.

3.8. Simulation setup

Sections 3.2–3.7 contain the theoretical framework. Our next task is to evaluate the performance of the WALS predictor in a number of common situations and in comparison with other often-used predictors. In the current section we describe the setup of our simulation experiment. The simulation results are presented in Section 3.9. Many extensions of the benchmark setup were considered and some of these are summarized in Sections 3.10 and 3.11.

3.8.1. Five methods

In the simulations we compare the performance of the WALS predictor to four commonly-used methods: unrestricted maximum likelihood (ML), pretesting (PT), ridge regression

(Ridge), and Mallows model averaging (MMA). We briefly describe each method below.

Unrestricted maximum likelihood simply estimates the unrestricted model (with *all* auxiliary regressors). There is no model selection here, and hence no noise associated with the model selection procedure. On the other hand, the noise associated with the estimation procedure will be large because of the large number of parameters.

Pretest estimation is a long-standing practice in applied econometrics, perhaps because pretest estimators are 'logical outcomes of the increased diagnostic testing of assumptions advocated in many econometric circles' (Poirier, 1995, p. 522). Pretest estimators and predictors do not follow textbook OLS or GLS properties, because the reported predictor is biased and its variance is only correct *conditional* on the selected model. One would expect the unconditional ('true') variance to be larger, because of the model selection noise. Giles and Giles (1993) provide a comprehensive review of the pretest literature. In pretest prediction one first selects the model based on diagnostic testing, and then predicts under the selected model. The choice of critical values of the pretest has received much attention (Toyoda and Wallace, 1976; Ohtani and Toyoda, 1980; Wan and Zou, 2003). Here we use the *stepwisefit* routine in Matlab, one of the most popular pretest methods. This routine begins with a forward selection procedure based on an initial model, then employs backward selection to remove variables. The steps are repeated until no additions or deletions of variables are indicated. We treat the model that includes only the focus regressors as the initial model and let the routine select the auxiliary regressors according to statistical significance. We choose the significance level for adding a variable to be 0.05 and for removing a variable to be 0.10.

Ridge regression (Hoerl and Kennard, 1970) is a common shrinkage technique, originally designed to address multicollinearity. Since the focus parameters are always in the model, we only penalize the auxiliary parameters. The ridge estimator is then obtained by minimizing

$$\phi(\beta_1, \beta_2) = (y - X_1\beta_1 - X_2\beta_2)'(y - X_1\beta_1 - X_2\beta_2) + \kappa\beta_2'\beta_2. \tag{3.43}$$

Letting

$$E_1 = \begin{pmatrix} I_{k_1} & 0_{k_1\times k_2} \\ 0_{k_2\times k_1} & 0_{k_2\times k_2} \end{pmatrix}, \qquad E_2 = \begin{pmatrix} 0_{k_1\times k_1} & 0_{k_1\times k_2} \\ 0_{k_2\times k_1} & I_{k_2} \end{pmatrix}, \tag{3.44}$$

the solution can be written as

$$\hat{\beta}(\kappa) = (X'X + \kappa E_2)^{-1}X'y, \tag{3.45}$$

where κ is the tuning parameter. Alternatively we obtain the ridge estimator in a Bayesian framework as the mean in the posterior distribution of $\beta|(X'X)^{-1}X'y$ by combining the data density $(X'X)^{-1}X'y|\beta \sim \mathrm{N}(\beta, \sigma^2(X'X)^{-1})$ with the partially informative prior $\beta/\sigma \sim \mathrm{N}(0, (1/\epsilon)E_1 + (1/\kappa)E_2)$ and letting $\epsilon \to 0$. Following Golub et al. (1979), we choose the tuning parameter κ by minimizing the generalized cross validation criterion

$$\mathrm{GCV}(\kappa) = \frac{(y - \Xi(\kappa)y)'(y - \Xi(\kappa)y)}{(N - \operatorname{tr}\Xi(\kappa))^2}, \qquad \Xi(\kappa) = X\left(X'X + \kappa E_2\right)^{-1} X'. \tag{3.46}$$

Finally, Mallows model averaging, proposed by Hansen (2007), averages over estimators using weights obtained by minimizing the Mallows criterion

$$C(\lambda) = (y - P(\lambda)y)'(y - P(\lambda)y) + 2\sigma^2 \operatorname{tr} P(\lambda), \tag{3.47}$$

where $\lambda = (\lambda_1, \ldots, \lambda_{2^{k_2}})$, $P(\lambda) = \sum_i \lambda_i X^{(i)}(X^{(i)\prime}X^{(i)})^{-1}X^{(i)\prime}$, and $X^{(i)}$ is the regressor matrix in model $\mathcal{M}_i$. Note that we do not assume an explicit ordering of the regressors, as Hansen does. An explicit ordering has the computational advantage that it reduces the number of weights from 2^{k_2} to k_2, but it is typically not practical in applications. When the submodels are strictly nested, Hansen (2007) proved that the MMA estimator is asymptotically optimal in a given class of model averaging estimators. Wan et al. (2010) extended the optimality to non-nested models.

All predictors explicitly account for possible correlation in the random disturbances. In particular, the WALS predictor is obtained using Definition 1, and the predictors of the other four predictors are all computed from

$$\hat{y}_f = X_f\hat{\beta} + C_f\Omega^{-1}(y - X\hat{\beta}), \tag{3.48}$$

where $\hat{\beta}$ depends on the chosen method. For ML (unrestricted model, no model selection), the predictor is linear in y and the associated variance is easily computed. For PT and Ridge, the predictor is not linear in y, but the reported variance is calculated as if the predictor were linear in y, following common practice. The variance for WALS is estimated from (3.39) while the variance for MMA can not be computed.

3.8.2. Data-generation process

We generate the data in three steps. First, we design the regressor matrix $X = (X_1 : X_2 : X_3)$, where X_1 and X_2 contain the focus and auxiliary variables, while X_3 contains the regressors that are omitted by the researcher (from *every* model) either because of ignorance or because of data limitations. The DGP and the largest (unrestricted) model are therefore not necessarily the same in the simulations. This is important because it brings us one step closer to econometric practice. In the benchmark DGP we consider six regressors with $k_1 = 2$, $k_2 = 3$, and $k_3 = 1$, such that

$$X_1 = (x_1, x_2), \qquad X_2 = (x_3, x_4, x_5) \qquad X_3 = (x_6), \tag{3.49}$$

where x_1 is the intercept. Since $k_2 = 3$ we have $2^3 = 8$ possible models. In the benchmark, all regressors, except the intercept, are generated by independent standard-normal distributions, and they are treated as fixed, so that each replication uses the same realization of the regressors once they have been generated. We also consider the cases where regressors are correlated (multivariate normal distribution with non-zero a covariance matrix) as a robustness check, and we find the results largely similar except that WALS always outperform other methods when there are omitted variables. In Section 3.11 we shall consider extensions where we have a large number of regressors and the regressors are autocorrelated or non-normally distributed.

Next, we simulate the parameters β_j $(j = 1, \ldots, 6)$ corresponding to regressors $x_1, \ldots, x_6$. For the auxiliary and omitted regressors $x_3, \ldots, x_6$ we set these parameters indirectly by controlling the 'theoretical' t-ratios, as follows. If we estimate the focus variables and just one auxiliary variable x_j, we obtain an estimated coefficient $\hat{\beta}_j$ with variance $\text{var}(\hat{\beta}_j) = (x_j' M_1^* x_j)^{-1}$. This implies a t-ratio $\hat{t}_j = \hat{\beta}_j \sqrt{x_j' M_1^* x_j}$. The 'theoretical' t-ratio is now defined as

$$t_j = \beta_j \sqrt{x_j' M_1^* x_j} \qquad (j = 3, \ldots, 6). \tag{3.50}$$

The values of the t_j are important (especially whether $|t_j| > 1$ or $|t_j| < 1$), because they determine whether adding an auxiliary regressor to the model will increase or decrease the root mean squared prediction error (the square root of the mean squared prediction error); see Theorem 1. We consider five combinations, as follows:

	Auxiliary			Omitted
T	t_3	t_4	t_5	t_6
T_1	1.2	0.9	1.1	0.0
T_2	1.2	1.7	0.7	0.9
T_3	1.2	0.9	1.0	2.5
T_4	2.0	2.5	2.7	0.0
T_5	0.4	0.2	0.5	0.0

Given x_j and t_j, we then obtain the parameters β_j $(j = 3, \ldots, 6)$. Three of the five cases (T_1, T_4, T_5) have no omitted variables. In T_1 the t-ratios of the auxiliary variables are close to 1, in T_4 the t-ratios are large, and in T_5 they are small. The other two cases (T_2, T_3) have an omitted variable. The value of t_6 is either close to one (T_2) or large (T_3).

Regarding the focus parameters we let $\beta_1 = \beta_2 = \nu\sqrt{\sum_{j=3}^{6}\beta_j^2}$ for three values of ν: 1, 2, and 3. Since the prediction performance is hardly affected by this choice, we shall report for $\nu = 2$ only.

Finally, we generate the error terms, based on (3.3), from a normal distribution with mean zero and variance Ω_{all}. We consider three specifications of Ω_{all}: homoskedasticity, heteroskedasticity, and autocorrelation. More precisely,

- homoskedasticity: $\Omega_{all} = \sigma^2 I_{n+n_f}$ with $\sigma^2 \in \{0.25, 1.00\}$;
- heteroskedasticity: $\Omega_{all} = \mathrm{diag}\left[\exp(\tau x_2)\right]$ with $\tau \in \{0.2, 0.7\}$;
- autocorrelation: AR(1) with $\sigma^2 = 1.0$ and $\rho \in \{0.3, 0.8\}$.

3.8.3. Comparison of prediction methods

We evaluate the five methods by comparing the predictors and the estimated variances of the predictors. To compare the predictors produced by the five methods, we consider the deviation between the predictor $\hat{y}_f$ and the true value y_f. A direct comparison is, however, misleading because there is a component common to all procedures. Hence we compute a modified version of the root mean squared prediction error,

$$\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{y}_f^{(r)} - y_f^{(r)} + (u_f - C_f\Omega^{-1}u)\right)'\left(\hat{y}_f^{(r)} - y_f^{(r)} + (u_f - C_f\Omega^{-1}u)\right)} \tag{3.51}$$

where $\hat{y}_f^{(r)}$ and $y_f^{(r)}$ are the predictor and the true value in the r-th replication. We follow Hansen (2008) and subtract $u_f - C_f\Omega^{-1}u$ from the prediction error, because it is common across prediction methods and independent of u, hence independent of $\hat{\beta} - \beta$.

To compare the prediction variances is more subtle. We could just compare the magnitudes of

$$\frac{1}{R}\sum_{r=1}^{R} \mathrm{var}(\hat{y}_f^{(r)}), \tag{3.52}$$

which would tell us whether one method reports more precise predictions than another. This is of interest, but more important than whether the reported prediction variance is small is whether the prediction variance is *correct*. It is easy to find predictors with small variances, but this does not make them good predictors.

Thus we wish to determine how close the estimated variance is to the 'true' variance, and this is measured by the RMSE of the prediction variance,

$$\sqrt{\frac{1}{R}\sum_{r=1}^{R}\Big(\mathrm{var}(\hat{y}_f^{(r)}) - V_T\Big)^2}, \tag{3.53}$$

where V_T denotes the 'true' variance, that is, the actual variance of the predictor. Since different methods give different predictors, the 'true' variance of the predictor varies across methods. We estimate V_T by obtaining $R_v = 100$ predictors from the replications, and then computing the sample variance of these predictors,

$$V_T := \frac{1}{R_v - 1}\sum_{r=1}^{R_v}\Big(\hat{y}_f^{(r)} - \frac{1}{R_v}\sum_{r=1}^{R_v}\hat{y}_f^{(r)}\Big)^2. \tag{3.54}$$

We consider training samples of size $N = 100$ and $N = 300$, and a prediction sample of size $N_f = 10$. The simulation results are obtained by computing averages across $R = 3000$ draws.

3.9. Simulation results: The benchmark

We compare the predictors by considering two sample sizes ($N = 100$, $N = 300$), five sets of parameter values ($T_1, \ldots T_5$), six specifications of Ω_{all}, and five methods. Each method is presented relative to ML, that is, we present the RMSE of each method divided by the RMSE of ML. An entry greater than one thus indicates an inferior performance relative to the ML method.

Table 3.1: RMSE of predictor relative to ML, benchmark model

N	T	WALS	PT	Ridge	MMA	WALS	PT	Ridge	MMA
					Homoskedasticity				
		$\sigma^2 = 0.25$				$\sigma^2 = 1.00$			
	T_1	**0.8644**	1.0570	0.9109	0.9416	**0.8981**	1.1589	1.0641	1.0172
	T_2	**0.8819**	1.0311	0.8979	0.9406	**0.9287**	1.1373	1.0350	1.0178
100	T_3	**0.9525**	1.2025	1.0866	1.0769	**0.9366**	1.1001	0.9576	1.0031
	T_4	1.0024	1.2756	1.0296	1.1579	**0.9569**	1.2758	1.0521	1.1267
	T_5	0.8190	0.8600	**0.7556**	0.8019	0.7939	0.8280	**0.6906**	0.7680
	T_1	**0.8990**	1.1152	0.9899	0.9940	**0.9295**	1.0421	0.9398	0.9727
	T_2	**0.9598**	1.0819	1.0237	1.0156	**0.9271**	1.0758	0.9648	0.9934
300	T_3	**0.9523**	1.0341	0.9612	0.9841	**0.9642**	1.0480	0.9929	0.9977
	T_4	**0.9516**	1.1490	0.9977	1.0853	1.0513	1.2362	1.0481	1.1492
	T_5	0.8004	0.8571	**0.7195**	0.7878	0.8404	0.8777	**0.7823**	0.8238
					Heteroskedasticity				
		$\tau = 0.2$				$\tau = 0.7$			
	T_1	**0.9038**	1.0766	0.9401	0.9708	**0.8926**	1.0964	0.9786	0.9815
	T_2	**0.8846**	1.1140	0.9936	0.9847	**0.8743**	1.0639	0.8699	0.9418
100	T_3	**0.9736**	1.0682	1.0071	1.0169	**0.8752**	0.9326	0.8681	0.8849
	T_4	**0.9628**	1.2428	0.9650	1.1144	**0.9558**	1.2430	0.9679	1.1158
	T_5	0.8160	0.8633	**0.7494**	0.7976	0.8148	0.8633	**0.7486**	0.7948
	T_1	**0.8928**	1.1104	0.9847	0.9917	**0.8834**	1.0718	0.8988	0.9581
	T_2	**0.9012**	1.0801	0.9810	0.9744	**0.8675**	1.0014	0.8694	0.9134
300	T_3	0.8884	0.9702	**0.8807**	0.9069	**0.9421**	1.0325	0.9530	0.9754
	T_4	1.0596	1.2178	1.0485	1.1456	**0.9011**	1.2004	0.9741	1.1056
	T_5	0.8043	0.8431	**0.7209**	0.7806	0.9132	0.9360	**0.8847**	0.9058
					Autocorrelation				
		$\rho = 0.3$				$\rho = 0.8$			
	T_1	**0.8875**	1.0543	0.9384	0.9586	**0.9809**	1.0182	0.9903	0.9958
	T_2	**0.8980**	1.0885	0.9001	0.9643	0.9659	1.0058	**0.9608**	0.9774
100	T_3	0.8823	0.9723	**0.8398**	0.9110	**0.9990**	1.0232	1.0126	1.0111
	T_4	1.0089	1.1629	1.0127	1.0917	**0.9979**	1.0511	1.0109	1.0278
	T_5	0.8634	0.8971	**0.8077**	0.8510	0.9632	0.9715	**0.9525**	0.9600
	T_1	**0.9188**	1.0868	0.9904	0.9934	**0.9760**	1.0241	0.9891	0.9947
	T_2	**0.9399**	1.0771	0.9975	0.9926	**0.9574**	0.9988	0.9603	0.9717
300	T_3	0.9031	0.9819	**0.8991**	0.9243	**0.9856**	1.0164	0.9920	0.9975
	T_4	1.0460	1.1804	1.0423	1.1207	**0.9988**	1.0623	1.0060	1.0373
	T_5	0.8453	0.8735	**0.7829**	0.8265	0.9788	0.9828	**0.9711**	0.9764

The RMSEs of the predictors are presented in Table 3.1. WALS comes out best in 39 out of 60 cases (65%), followed by Ridge (27%) and ML (8%). The pretest and MMA predictors never dominate. The dominance of WALS occurs for each of the specifications of Ω_{all}, though slightly less in the autocorrelation case than in the homo- and heteroskedastic cases. One reason why WALS is superior over MMA is that WALS makes use of the information in the error structure, while MMA does not.

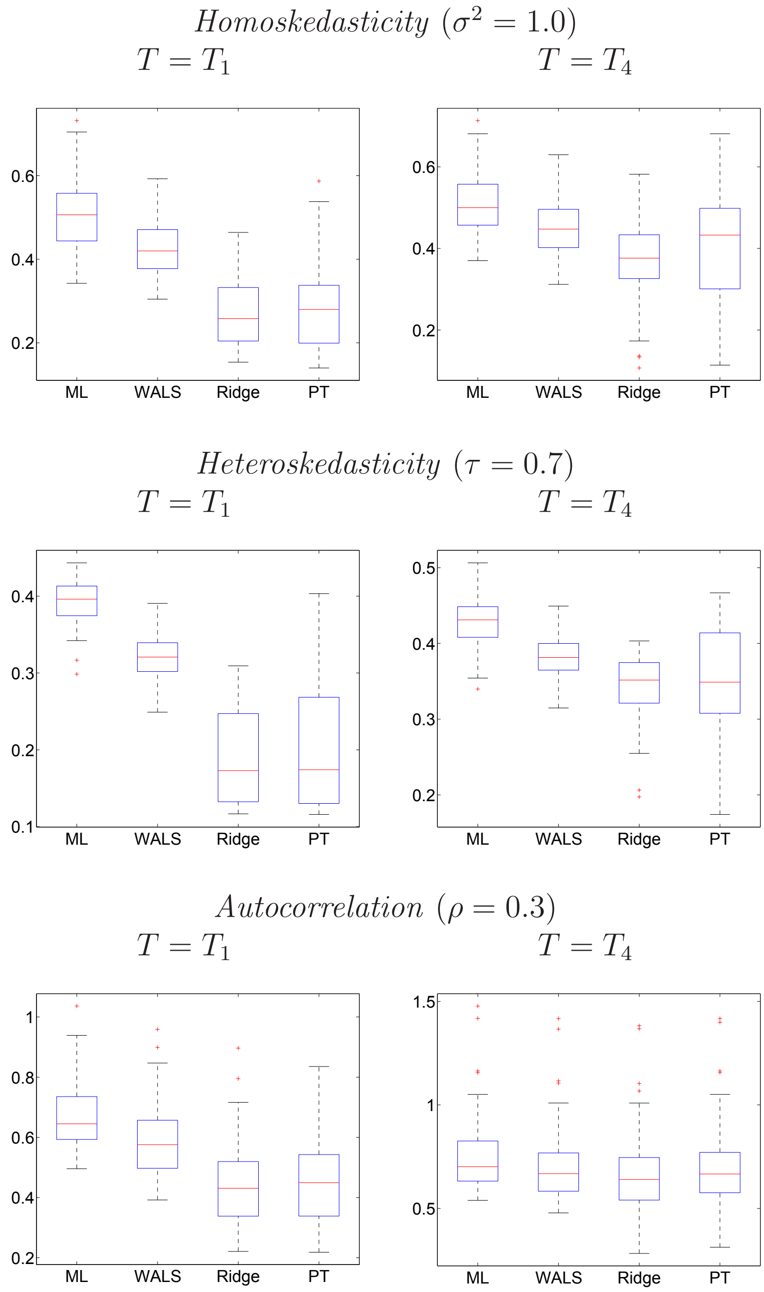
In T_1 WALS dominates in all 12 cases, and in T_2 in 11/12 cases. This shows that WALS performs well when the t-ratios of the auxiliary variables are close to one, even when the model possibly omits one variable with a t-ratio close to one. If the omitted variable has a stronger impact on the dependent variable, as in T_3, WALS still works best in 9/12 cases followed by Ridge (3/12). This suggests that omitting important regressors may affect the prediction ability of WALS, which is a point worthy of further investigation; see Section 3.11.

When the t-ratios of the auxiliary variables are much larger than one, as in T_4, then WALS is still the best, but this is the only case where ML also performs well. This makes sense, because model uncertainty plays a smaller role here. We note that increasing the parameter values in Ω_{all} (σ^2, τ, and ρ, respectively) improves the relative performance of WALS over ML in T_4. Larger parameter values in Ω_{all} imply more noise in the model, and the superiority of WALS seems stronger then. The possibility that the degree of model uncertainty affects the relative performance of different methods also warrants further investigation; see Section 3.10.

In the opposite case where the t-ratios of the auxiliary variables are much smaller than one, as in T_5, WALS is not the best. Here the Ridge predictor always dominates, and ML is always the worst. Again, there is little model uncertainty. The unrestricted model (ML) is not appropriate, but shrinkage towards the restricted estimator (with only the focus regressors) makes sense, and this is what Ridge does. We also experiment the cases with negative t-ratios, and find that the results are not substantially different. WALS comes out best in most cases (28%), followed by ridge (15%).

Next we compare the performance of the prediction variance. We first consider the magnitude of the estimated variance itself, then we ask how close the estimated variance is to the 'true' variance. The MMA method is not included in this comparison because there is no procedure known to us to compute this variance. In the boxplots of Figure 3.1, the central mark is the median, the edges of each box indicate the 25-th and 75-th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

Figure 3.1: Estimated variance in the benchmark model ($N = 100$)



We consider six representative cases. Judging by the median of the estimated variance, ML has the largest variance, followed by WALS, while the variance of the Ridge and PT predictors are both smaller than WALS. This is in accordance with intuition, because ML includes all regressors, while pretesting and ridge are based on the selected model or the selected parameter, while ignoring variation caused by the selection procedure. The WALS predictor has a relatively large variance (but still smaller than ML), because it does take the uncertainty in the selection procedure into account.

We note that the estimated variances for WALS and ML are more concentrated on their median values than those of Ridge and PT, and that the distributions of the latter two methods are also characterized by a strong asymmetry. The difference between the four variance estimates is relatively small when there is little model uncertainty (T_4), and more pronounced when model uncertainty is large (T_1).

As discussed above, a variance estimate is a good estimate, not when it is small, but when it provides the correct information about the precision of the predictor. If this precision happens to be low, then we need to provide a high value for the variance estimate. Table 3.2 gives the RMSE of the estimated prediction variance, as given in (3.53), again relative to ML. On the left-side of the table (where the parameters σ^2, τ, and ρ are relatively small), the RMSE ratios (relative to ML) are, on average, 1.10 for WALS, 2.43 for Ridge, and 10.98 for PT. On the right-side (where the parameter values are larger, hence more uncertainty), the RMSE ratios are 1.05 for WALS, 2.19 for Ridge, and 9.44 for PT. The main conclusion from the table is therefore that ML and WALS provide the best estimates of the prediction variance, while Ridge and especially PT generally report a variance which is misleadingly small. While WALS provides a much better estimate of the forecast than ML, the variance of the forecast is slightly more accurately estimated in ML than in WALS.

ML performs particularly well when N is large (because of the asymptotic behavior of ML estimates and predictions) and when the variance parameters are small. The relative performance of WALS prediction variance estimates is improved by increasing the variance of the error terms. This suggests that more model uncertainty makes WALS prediction more attractive. In the benchmark setup, where we have assumed deterministic regressors and coefficients, there is not much model uncertainty. If we raise the model uncertainty, for example by introducing random regressors or random coefficients or by increasing the variance of the errors, then one would expect the WALS estimates, which incorporate the model uncertainty, to be more accurate than ML. We shall analyze this idea further in the next section.

Table 3.2: RMSE of prediction variance relative to ML, benchmark model

N	T	WALS	PT	Ridge	WALS	PT	Ridge
		Homoskedasticity					
		$\sigma^2 = 0.25$			$\sigma^2 = 1.00$		
	T_1	**0.8155**	10.417	2.1262	**0.8050**	9.7509	2.1211
	T_2	**0.9011**	12.307	2.8462	**0.8606**	11.722	2.7199
100	T_3	1.0760	8.5721	1.9561	1.0421	9.7699	2.0451
	T_4	**0.7755**	16.040	2.1837	**0.7972**	16.747	2.2023
	T_5	0.7765	3.0180	**0.6894**	0.7916	3.0839	**0.7061**
	T_1	1.2139	16.351	3.3701	1.2947	15.494	3.0947
	T_2	1.1475	16.967	3.7704	1.3978	18.440	4.0967
300	T_3	1.3510	15.950	3.3156	1.4509	15.176	3.3333
	T_4	1.0001	21.714	2.5714	**0.8938**	20.070	2.5663
	T_5	1.4811	5.3371	1.0851	1.2495	4.9051	1.0459
		Heteroskedasticity					
		$\tau = 0.2$			$\tau = 0.7$		
	T_1	1.1174	17.938	4.6738	**0.8849**	12.547	3.3035
	T_2	1.0030	23.113	5.3124	**0.9509**	18.382	4.2944
100	T_3	1.1227	25.641	5.3652	**0.9648**	18.941	4.8476
	T_4	1.0083	17.748	3.3204	**0.9628**	15.244	2.8077
	T_5	1.2374	5.1309	1.6143	**0.9888**	4.5029	1.5134
	T_1	1.3897	18.963	3.8448	1.1613	16.107	3.3285
	T_2	1.3279	20.426	4.4850	1.2645	17.335	4.3203
300	T_3	1.4954	19.552	3.9925	1.3081	16.980	3.2964
	T_4	1.2030	27.566	3.4499	1.0624	22.479	2.7482
	T_5	1.1378	5.2663	1.1331	1.2631	5.2922	1.2149
		Autocorrelation					
		$\rho = 0.3$			$\rho = 0.8$		
	T_1	1.0605	2.5141	1.2681	1.0000	1.0156	1.0064
	T_2	1.0347	2.9666	1.3984	1.0011	1.0141	1.0072
100	T_3	1.0381	2.8648	1.3693	**0.9984**	1.0213	1.0093
	T_4	1.0190	3.6270	1.2758	1.0007	1.0162	1.0047
	T_5	1.0633	1.4584	1.0620	1.0000	1.0088	1.0049
	T_1	1.0060	1.6103	1.0716	**0.9983**	1.0050	1.0003
	T_2	1.0191	1.6745	1.1390	**0.9997**	1.0065	1.0016
300	T_3	1.0080	1.6066	1.0980	**0.9991**	1.0087	1.0016
	T_4	1.0068	1.7774	1.0753	1.0006	1.0031	1.0009
	T_5	1.0258	1.1822	1.0238	1.0000	1.0035	1.0019

3.10. Simulation results: More uncertainty

In this section we extend the benchmark setup by introducing additional randomness in the model. This is achieved by allowing for random regressors or random coefficients or by increasing the variance of errors.

3.10.1. Random regressors

We first consider the model with random but exogenous regressors. This is a common extension in simulation designs, and particularly useful in applications where one wishes to model dynamic economic behavior. The only difference with the benchmark is that we generate a new set of X's from $\mathrm{N}(0, \sigma_x^2)$ in every replication, so that each realization of the y-series involves a new realization of the X-series. (The introduction of σ_x^2 is unimportant, because the RMSE is invariant to its value.) The generation of X is independent of the errors.

Allowing the regressors to be random increases the RMSE of the forecast in each method. The relative performance of the five predictors is similar to the benchmark case. In particular, the WALS predictor has the lowest RMSE in T_1, T_2, and T_3, about 5% lower than the RMSE of the ML predictor. In case T_5, the ridge predictor has the lowest RMSE under all error structures, around 10% lower than the ML predictor. In contrast to the benchmark results, allowing random regressors improves the relative performance of WALS over ML in T_4, because more randomness decreases the importance of the auxiliary variables.

The main difference between the random regressor model and the benchmark model is in the prediction variance, and we report its RMSE in Table 3.3. WALS now produces the most accurate prediction variance in all cases, including T_4 and T_5. This remarkable performance of WALS is due to the fact that randomness in the regressors raises model uncertainty, which in turn increases the variation of the predictor, that is, the true variance. The prediction variance of WALS explicitly incorporates such model uncertainty, in contrast to pretesting, ridge regression, and ML.

Table 3.3: RMSE of prediction variance relative to ML, random regressor model

N	T	WALS	PT	Ridge	WALS	PT	Ridge
				Homoskedasticity			
		$\sigma^2 = 0.25$			$\sigma^2 = 1.00$		
	T_1	**0.7499**	1.0219	0.8056	**0.7467**	1.0119	0.8007
	T_2	**0.7958**	1.0096	0.8365	**0.7952**	1.0110	0.8362
100	T_3	**0.8866**	1.0101	0.9149	**0.8899**	1.0161	0.9220
	T_4	**0.8487**	0.9951	0.8990	**0.8497**	0.9929	0.8987
	T_5	**0.5091**	0.9267	0.5336	**0.5064**	0.9021	0.5229
	T_1	**0.7486**	1.0219	0.8064	**0.7435**	1.0165	0.8011
	T_2	**0.7978**	1.0107	0.8407	**0.7990**	1.0101	0.8399
300	T_3	**0.8930**	1.0170	0.9222	**0.8889**	1.0115	0.9179
	T_4	**0.8473**	0.9953	0.8989	**0.8474**	0.9938	0.8988
	T_5	**0.5123**	0.9513	0.5505	**0.5147**	0.9417	0.5481
				Heteroskedasticity			
		$\tau = 0.2$			$\tau = 0.7$		
	T_1	**0.7448**	1.0296	0.8095	**0.7421**	1.0342	0.8128
	T_2	**0.7914**	1.0123	0.8381	**0.7950**	1.0129	0.8448
100	T_3	**0.8862**	1.0080	0.9149	**0.8835**	1.0067	0.9137
	T_4	**0.8461**	0.9954	0.9010	**0.8495**	0.9925	0.9044
	T_5	**0.5125**	0.9515	0.5525	**0.5621**	0.9278	0.5855
	T_1	**0.7444**	1.0192	0.8054	**0.7444**	1.0192	0.8054
	T_2	**0.7987**	1.0162	0.8450	**0.7987**	1.0162	0.8450
300	T_3	**0.8906**	1.0139	0.9191	**0.8906**	1.0139	0.9191
	T_4	**0.8490**	0.9942	0.9014	**0.8490**	0.9942	0.9014
	T_5	**0.5146**	0.9614	0.5548	**0.5146**	0.9614	0.5548
				Autocorrelation			
		$\rho = 0.3$			$\rho = 0.8$		
	T_1	**0.7469**	1.0318	0.8150	**0.9809**	1.0059	0.9890
	T_2	**0.7997**	1.0241	0.8494	**0.9683**	1.0049	0.9785
100	T_3	**0.8866**	1.0175	0.9177	**0.9646**	1.0065	0.9775
	T_4	**0.8479**	0.9986	0.9022	**0.9099**	1.0004	0.9433
	T_5	**0.6863**	0.9902	0.7205	**0.9975**	1.0082	1.0033
	T_1	**0.7629**	1.0151	0.8199	**0.9924**	1.0003	0.9950
	T_2	**0.8054**	1.0122	0.8483	**0.9894**	1.0058	0.9940
300	T_3	**0.8913**	1.0082	0.9201	**0.9848**	0.9996	0.9898
	T_4	**0.8478**	0.9934	0.9005	**0.9465**	1.0023	0.9666
	T_5	**0.8013**	0.9948	0.8239	0.9985	0.9990	**0.9977**

3.10.2. Random coefficients

Next we consider the situation where the coefficients of the explanatory variables are subject to random variation, that is,

$$y_t = \sum_{j=1}^{6} x_{tj}(\beta_j + v_{tj}) + u_t \qquad (t = 1, 2, \ldots, N), \tag{3.55}$$

where the v_{tj}'s are independent unobserved random disturbances, distributed as $\mathrm{N}(0, \sigma_v^2)$. Such models date back to Rubin (1950), Hildreth and Houck (1968), Swamy (1970), Froehlich (1973), and others, who discussed parameter estimation and provided empirical applications. Prediction in random coefficient models is studied, *inter alia*, in Bondeson (1990) and Beran (1995). We can rewrite (3.55) as

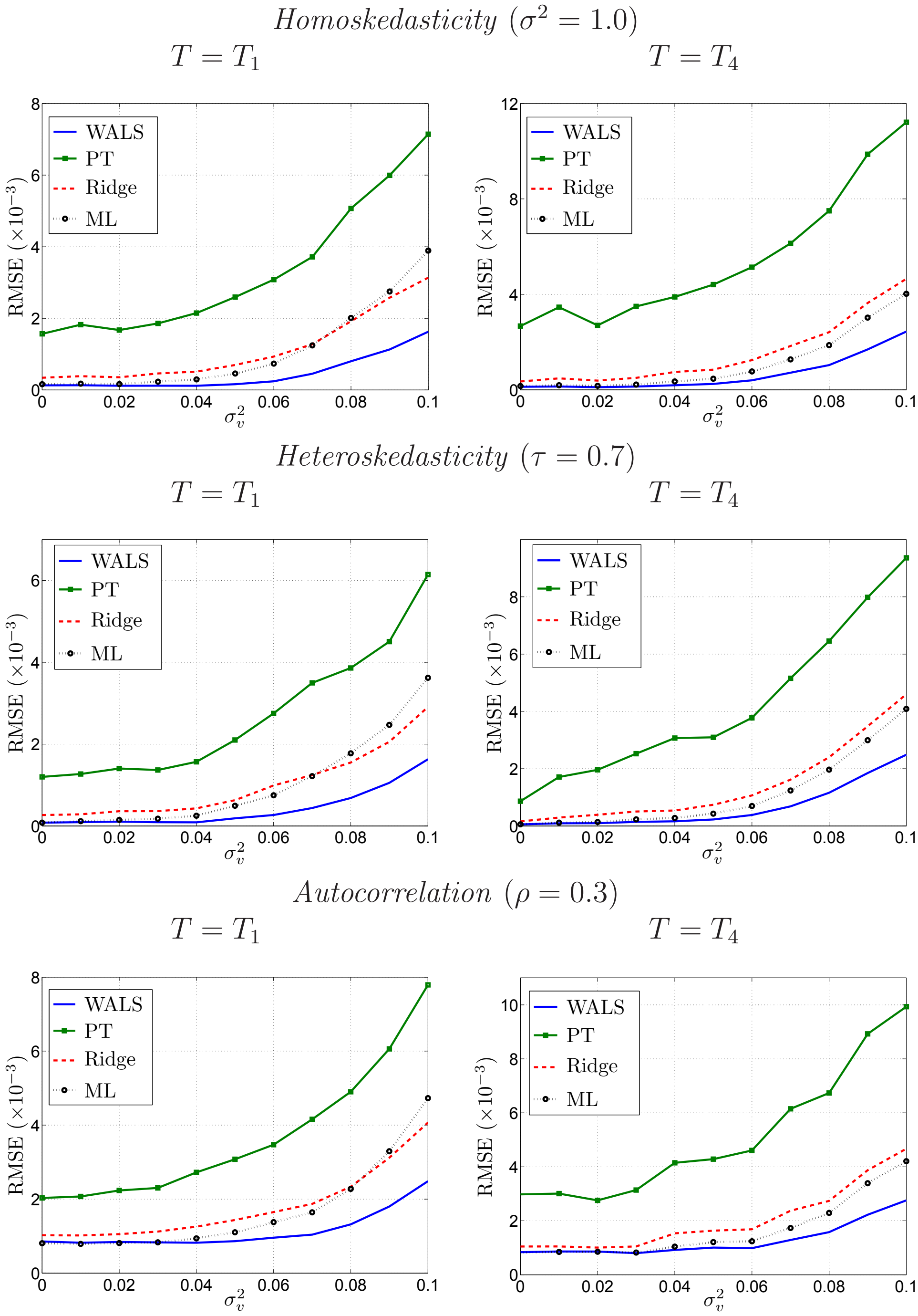
$$y_t = \sum_{j=1}^{6} x_{tj}\beta_j + \zeta_t, \qquad \zeta_t = \sum_{j=1}^{6} x_{tj}v_{tj} + u_t \tag{3.56}$$

where ζ_t is normally distributed with mean zero and variance $\sigma_\zeta^2 = \sigma_u^2 + \sigma_v^2 \sum_j x_{tj}^2$. This shows that introducing variation in the coefficients increases the variance of the errors. We assume that the researcher is ignorant of the random coefficients and misspecifies them as fixed. Hence the model is the benchmark model, but the DGP has changed. How do the predictors respond to this situation?

Regarding the accuracy of the predictors, we find similar results as in the random regressor model. The WALS predictor has the lowest RMSE in cases T_1–T_4, while the ridge predictor is the best under T_5. This demonstrates good performance of the WALS predictor when the t-ratios of the auxiliary variables are close to one, even when the coefficients are misspecified.

The accuracy of the estimated prediction variance is shown in Figure 3.2 as a function of σ_v^2. Increasing σ_v^2 raises the model uncertainty as well as the degree of misspecification, thus lowering the accuracy of all predictions. The variance estimates obtained from pretesting have a much larger RMSE than those from other methods, and they are also more volatile. Ridge regression generally produces somewhat better variance estimates. Most accurate are ML and WALS, and their variance accuracy is close when σ_v^2 is small. When $\sigma_v^2 = 0$ (the benchmark), ML is more accurate than WALS, but as σ_v^2 increases, the RMSE of WALS increases slower than the RMSE of ML, and when $\sigma_v^2 > 0.03$ the accuracy of WALS variance estimates is higher than ML. These results confirm that

Figure 3.2: RMSE of prediction variance in random coefficient model ($N = 100$)



WALS behaves well in the presence of a large degree of model uncertainty. Viewed differently, WALS is more robust than pretesting, ridge, and ML.

3.10.3. Increase in the variance of errors

Finally, we consider an increase in the variance of the errors by changing a parameter in Ω_{all}. We only consider the homoskedastic and the heteroskedastic cases. Under homoskedasticity we can increase the error variance by increasing σ^2; under heteroskedasticity case by increasing τ.

Figure 3.3: RMSE of prediction variance: homoskedastic versus heteroskedastic ($N = 100$, $T = T_1$)

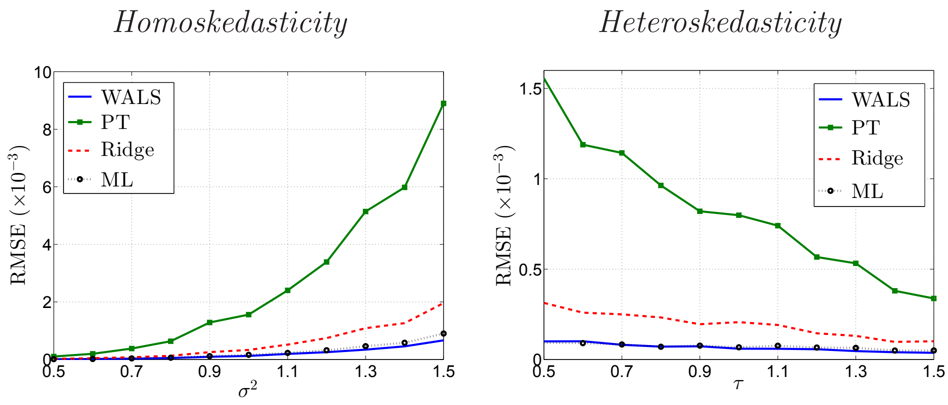


Figure 3.3 shows how the RMSE of the prediction variance changes as the parameters σ^2 and τ increase. In both cases, WALS and ML outperform Ridge and, in particular, PT. When the error variance is small, the prediction variances produced by WALS and ML show similar accuracy. But as the error variance increases, the WALS prediction variance is more accurate than ML.

Note that increasing the error variance affects the RMSE of the prediction variance in different ways: it increases the RMSE in the homoskedastic case but reduces the RMSE in the heteroskedastic case. This is because in the design of the heteroskedastic variance, $\Omega_{all} = \exp(\tau x_2)$ is a function of x_2. Increasing τ leads to a smaller estimated coefficient $\hat{\beta}_2$ since the estimation process cannot distinguish between increasing the error variance from increasing the variation in x_2.

In summary, more model uncertainty leads to a better performance of WALS relative to the other methods.

3.11. Simulation results: More regressors

In Sections 3.9 and 3.10 we assumed two focus regressors, three auxiliary regressors, and one omitted regressor. In practical applications the number of regressors is likely to be larger. In this section we extend the benchmark framework by assuming $k_2 = 12$ auxiliary regressors and $k_3 = 3$ omitted regressors, while keeping the same number $k_1 = 2$ of focus regressors. The large number of auxiliary regressors will increase the model uncertainty, because we now have $2^{12} = 4096$ different models to consider compared to $2^3 = 8$ in the benchmark. When introducing new variables we have to specify the 'theoretical' t-ratios which are used to compute the values of the β-parameters. We consider four combinations, as follows:

	Auxiliary	Omitted
T	t_3–t_{14}	t_{15}–t_{17}
T_{L1}	1.2, 0.9, 1.0, 1.3, 1.2, 1.5, 1.6, 1.2, 1.1, 0.8, 1.5, 1.4	0.0, 0.0, 0.0
T_{L2}	1.2, 0.9, 1.0, 1.3, 1.2, 1.5, 1.6, 1.2, 1.1, 0.8, 1.5, 1.4	2.4, 2.8, 2.0
T_{L3}	1.2, 0.9, 1.0, 2.3, 2.2, 2.5, 2.6, 2.1, 2.0, 0.5, 2.5, 1.4	0.0, 0.0, 0.0
T_{L4}	1.2, 0.9, 1.0, 0.7, 1.2, 0.5, 0.6, 2.2, 0.3, 0.8, 0.5, 1.2	0.0, 0.0, 0.0

In T_{L1} all auxiliary variables have t-ratios close to one and there are no omitted variables. In T_{L2} we have the same t-ratios for the auxiliary variables but now there are also omitted variables. In T_{L3} many of the auxiliary variables have 'large' t-ratios, while in T_{L4} many of the t-ratios are 'small'. Only T_{L2} has omitted variables and they are all important. We combine this larger data set with the benchmark setup, random regressor DGP, and random coefficient DGP, again under each of the three error structures. We compare WALS, Ridge, and PT with ML. We do not compute MMA because the computational burden is too high when k_2 is large.

We briefly consider two other extensions, both analyzed in the context of the small data set: dependence among the regressors and non-normality. Dependence is introduced through a multivariate normal distribution with correlation 0.5, while the non-normal regressors are obtained from a Student distribution with five degrees of freedom. We also considered correlated regressors, e.g. AR(1) regressors. These results are essentially the same and therefore not reported. We experiment (separately) with these two extensions in the benchmark model and also in models with more uncertainty. The simulation results are largely similar to the case with normal and uncorrelated regressors. In particular, the WALS predictor is the most accurate when t-ratios are close to one, and the WALS prediction variance is particularly reliable when there is additional uncertainty.

Table 3.4: RMSE relative to ML, many auxiliary regressors ($N = 100$)

T	*Homoskedasticity* ($\sigma^2 = 1.0$) WALS	PT	Ridge	*Heteroskedasticity* ($\tau = 0.7$) WALS	PT	Ridge	*Autocorrelation* ($\rho = 0.3$) WALS	PT	Ridge
				Benchmark model: fixed X, fixed β					
	Predictor								
T_{L1}	**0.8611**	1.2088	0.8982	**0.8570**	1.1664	0.9411	**0.9284**	1.0807	0.9561
T_{L2}	0.9289	1.0789	**0.9234**	0.8755	1.0167	**0.8328**	0.9398	1.0637	**0.9229**
T_{L3}	**0.9625**	1.3057	0.9883	**0.9047**	1.2300	0.9294	**0.9052**	1.1423	0.9077
T_{L4}	0.8285	1.0579	**0.7997**	0.8035	0.9935	**0.7745**	0.8993	1.0058	**0.8760**
	Prediction variance								
T_{L1}	**0.3440**	14.835	0.8088	**0.7854**	22.311	1.6711	1.3175	13.641	1.4356
T_{L2}	1.1154	17.104	**0.9023**	1.1196	25.152	1.4803	1.2224	15.158	1.5576
T_{L3}	**0.4239**	20.146	0.9557	**0.6404**	27.417	1.2445	1.1870	16.359	1.2598
T_{L4}	**0.3452**	8.6851	0.6884	**0.6729**	12.559	1.0391	1.2984	7.9238	1.3222
				Random regressor model: random X, fixed β					
	Predictor								
T_{L1}	**0.9787**	1.0249	0.9791	0.9828	1.0136	**0.9823**	**0.9810**	1.0188	0.9821
T_{L2}	0.9769	1.0182	**0.9761**	0.9811	1.0106	**0.9801**	0.9778	1.0132	**0.9753**
T_{L3}	**0.9854**	1.0365	0.9874	**0.9891**	1.0249	0.9908	**0.9868**	1.0327	0.9876
T_{L4}	0.9745	1.0023	**0.9715**	0.9821	1.0000	**0.9800**	0.9770	1.0022	**0.9754**
	Prediction variance								
T_{L1}	0.7713	1.0240	**0.7595**	**0.7634**	1.0329	0.7830	**0.7650**	1.0538	0.7779
T_{L2}	0.8569	1.0177	**0.8489**	0.8491	1.0024	0.8433	0.8523	1.0167	**0.8438**
T_{L3}	**0.8324**	1.0074	0.8443	**0.8277**	1.0003	0.8568	**0.8253**	1.0184	0.8498
T_{L4}	0.7365	1.0358	**0.7195**	**0.7299**	1.0636	0.7516	**0.7303**	1.0899	0.7540
				Random coefficient model: fixed X, random β					
	Predictor								
T_{L1}	**0.9896**	1.0216	0.9960	0.9913	1.0050	**0.9907**	**0.9844**	1.0179	0.9876
T_{L2}	**0.9926**	1.0400	1.0166	**0.9747**	0.9978	0.9836	0.9683	1.0027	**0.9616**
T_{L3}	**0.9748**	1.0309	0.9775	**0.9941**	1.0203	0.9966	**0.9966**	1.0333	0.9988
T_{L4}	**0.9763**	1.0157	0.9893	**0.9846**	1.0067	0.9866	0.9839	1.0055	**0.9833**
	Prediction variance								
T_{L1}	**0.1869**	6.9005	0.5812	**0.1577**	7.6836	0.7459	**0.4234**	8.3586	0.9135
T_{L2}	**0.3387**	11.519	0.8175	**0.1607**	9.1310	0.8254	**0.3520**	8.2753	0.8903
T_{L3}	**0.2718**	8.3617	0.7382	**0.2199**	9.3498	0.8352	**0.4600**	10.759	0.9792
T_{L4}	**0.1638**	4.2507	0.4721	**0.1532**	5.1157	0.6335	**0.4300**	5.7437	0.8624

3.12. Conclusion

This paper has introduced a new method of prediction averaging using weighted average least squares (WALS). We have argued that pretesting—the currently dominant prediction method—is dangerous, because it ignores the noise associated with model selection. Indeed, our simulation results demonstrate that pretesting performs very badly. Model averaging is an attractive method in that it allows us to combine model selection and prediction into one procedure. Within the model averaging methods we proposed the WALS predictor and also an estimate for its variance. Our predictor explicitly allows for correlation in the errors.

We have compared the WALS predictor with four competing predictors (unrestricted ML, pretesting, ridge regression, Mallows model averaging) in a wide range of simulation experiments, where we considered not only the accuracy of the predictor (measured by the root mean squared prediction error), but also the accuracy of the prediction variance. The WALS predictor generally produces the lowest mean squared error. The estimated variance of the WALS predictor, while typically larger than the variance of the pretesting and ridge predictors, is more accurate, and when model uncertainty increases the dominance of WALS becomes more pronounced. These results, together with the fact that the WALS predictor is easy to compute, suggest that the WALS predictor is a serious candidate in economic prediction and forecasting.

Chapter 4

Natural resource, institutional quality, and economic growth in China[3]

4.1. Introduction

Since reformists within the Chinese Communist Party initiated a program of economic reforms in December 1978, China has been the world's fastest-growing major economy with consistent growth rates of around 10% over the past thirty years. China is also the largest exporter and second largest importer of goods in the world. At the same time the production of natural resources has increased sharply. These natural resources are not evenly distributed over China: the coal mines are primarily located in eight provinces, all in the North-East and North, while most natural gas reserves can be found in the Mid-West, especially in Sichuan province which accounts for almost 30% of the nation's production of natural gas. Regions with a high production of natural resources have generally developed slower than low-producing regions, a phenomenon which resembles the situation where resource-rich countries perform worse than resource-scarce countries, the so-called 'curse of resources'.

The 'curse of resources' hypothesis has been analyzed in many cross-country studies, both from empirical and theoretical viewpoints, but there have not been many within-country studies examining the relationship between natural resources and economic growth. A notable exception is the study by Papyrakis and Gerlagh (2004), who employed data from 49 states in the USA, and concluded that resource-scarce states outperform resource-rich states. Like the USA, China is endowed with several unique characteristics which make it suitable for testing the resource curse hypothesis. First, China has homogeneous constitution, law, and governance structures (but different institutions) across provinces. Second, there are significant differences between provincial

[3]This chapter is coauthored with Jan R. Magnus and Kan Ji.

economies, and substantial variation in resource endowments and development. Third, market reforms have lifted restrictions on the flows of products, labor, and capital Zhang et al. (2008). In addition, the price reforms in China's natural resource sector between the late 1970s and the mid-1990s ensure that the resource prices largely reflect market supply and demand.

Recently, a number of studies have appeared on the relationship between resources and economic growth in China. Xu and Wang (2006) were the first to use panel-data methods, and they found evidence supporting the curse of resources at the provincial level. Shao and Qi (2009) confirmed these results and compared the resource effects before and after the 'West China Development Drive' in 2000, by estimating two samples (before and after the policy change) separately. Their results suggest that the 2000 policy change induced a resource curse. Zhang et al. (2008) employed a panel-data set at the provincial level and associated a slower growth rate of per capita consumption with rich resources, especially in rural regions. Fan et al. (2012) used city-level data to analyze the transmission mechanism of resource curse and diffusion processes of resources among cities. They found no evidence of a resource curse in China, and they showed that resources have a positive diffusion effect among neighboring cities within the same province.

Some caution is required in interpreting these results. First, no distinction is made between resource abundance and resource dependence. For example, Xu and Wang (2006) measured resource abundance by the proportion of mining workers or by the ratio of investment in the mining industry to total fixed asset investment. These measurements capture resource dependence rather than resource abundance, and the effect of these two concepts on economic growth is not necessarily the same (Brunnschweiler and Bulte, 2008). In addition, the measurement of resource dependence suffers from endogeneity (Brunnschweiler and Bulte, 2008; Norman, 2009; van der Ploeg and Poelhekke, 2010). The current paper measures resource abundance rather than resource dependence, thus avoiding these problems. Second, the analysis of the critical role of institutions in the association between resource abundance and economic growth is not satisfactory. Not only is the measurement of institutional quality poor, but also important nonlinearities are ignored (Ross, 2001). Finally, while panel-data methods capture short-run dynamics, they are typically not powerful in explaining the long-run effect of natural resources.

Conventional panel-data models estimate constant slope coefficients, implicitly assuming that the resource effect does not change over time. This may, however, not be the case in China, especially for regions with significant structural breaks such as the West China Development Drive.

In this paper, we study the interplay between resource abundance, institutional quality, and economic growth in China. We also investigate whether the resource effect on economic growth varies over time. Our paper makes four main contributions in the context of provincial China. First, we propose several new measurements of resource abundance. These new measurements consider resource abundance either as a stock or as a flow, thus allowing a comparison between *in situ* resource reserves (a stock) and resource revenues (a flow, usually referred to as a 'windfall gain'). Second, we re-examine the role of institutional quality in the relationship between resource abundance and economic growth. Institutional quality is proxied by confidence in the courts, using data from the World Bank. We investigate whether and how the effect of resource abundance on economic growth depends on institutional quality, employing a functional-coefficient model. Our results show that the effect of resource abundance in China depends on institutional quality in a nonlinear fashion, which can not be fully captured using a linear model. More importantly, we find — in contrast to Mehlum et al. (2006) — that the effect of natural resources is more positive for provinces with poor institutional quality. Third, we consider the West China Development Drive as a significant policy shock that may influence the effect of resource abundance on economic growth. We employ both a standard panel-data model and a time-varying coefficient model to study whether and how the resource effect changes after the policy shock. Finally, our paper uses both cross-section and panel data to explore the effect of natural resource abundance on economic growth. The advantage of cross-section data is that they better capture the long-run effect, and reduce the possible bias caused by economic fluctuations. The advantage of panel data is that they contain more information on the dynamics.

We employ provincial data over the period 1990–2008. Our results are not in general agreement with most of the current literature. First, the cross-section benchmark model shows no evidence of a resource curse. Second, the difference-in-difference approach shows that the interaction effect of resource abundance and institutional quality is positive but not significant, suggesting that the interaction effect may not be linear.

Third, extending the benchmark model, the functional-coefficient estimates indicate that resource abundance is strongly related to economic growth in regions where institutional quality is weak, and weakly related in regions where institutional quality is moderate. Fourth, both the standard panel-data approach and the time-variant model show that the West China Development Drive has had an important impact on the role of resources in the economy. By intensifying resource exploitation of the Western provinces, the Drive has led to an income rise in the West. This increased income helped the local economy for a short period, but not for long, possibly because an overemphasis on resource exploitation in some Western provinces crowded out other sectors to some extent.

The paper is organized as follows. In Section 4.2 we briefly review the theories relating resources and institutional quality, and formulate the questions raised in this paper. In Section 4.3 we describe the data, and present some characteristics and preliminary analysis. In Section 4.4 we present the cross-section analysis, and in Section 4.5 the panel-data analysis. Some conclusions are offered in Section 4.6.

4.2. Resources and institutional quality

Ever since the 1950s, economists have observed that resource-rich countries may grow slower than resource-scarce countries. Why do abundant resources tend to impede economic growth? Several theories have been developed, mainly Dutch disease models (Sachs and Warner, 1995) and institutional explanations. Traditional Dutch disease explanations cannot be directly applied in the Chinese context, because most of China's exports are not expensive for other countries to buy because labor is inexpensive in China. While it is possible to study the 'Dutch disease' among provinces, an adapted definition and appropriate data would be required. Due to the data limitations, we focus here on institutional quality explanations.

Many papers have stressed the importance of institutions through which abundant resources may curse economic growth. From a qualitative point of view, resource revenues appear to be easily appropriable, thus leading to rent-seeking behavior and corruption. Also, more labor is attracted to seek revenues from other productive activities (Isham et al., 2005; Leite and Weidmann, 1999; Norman, 2009). Auty (2001) argued that resource wealth promotes the ascendance of the 'predatory state' over the 'development state', either by encouraging the former through corruption or by undermining the latter when revenues associated with resource extraction reduce the efficiency of policy and administration. The relationship between resources and institutions also depends on the type of resources. Many studies show that 'point' (concentrated) resources result in poor institutions, while 'diffuse' resources do not. This is because point resources (such as oil, minerals, and plantations) are extracted from a narrow geographic or economic base, and can be protected and controlled at a relatively modest cost. In contrast, diffuse natural resources (such as agricultural products) are spread in space and utilized by agents characterized by horizontal relationships (Bulte et al., 2005). The latter are therefore less correlated with institutional quality.

From a quantitative point of view, Leite and Weidmann (1999) were perhaps the first to demonstrate the effect of resource abundance on institutional quality. Mehlum et al. (2006) interacted natural resource abundance with institutional quality and found that the negative effect of natural resources on economic growth only occurs in countries with poor institutional quality. Ross (2001) argued that institutions themselves may also be endogenous and not invariant with respect to resource endowments. Some empirical studies claim that institutional quality alone can explain a great deal of cross-country differences in economic development, thus further questioning the role of natural resources in economic development (Acemoglu et al., 2001).

The economy of China is in transition, hence it is a mixture of a market economy and a planned economy. This mixture is also reflected in the resource market. Before 1990 the Chinese central government controlled the price of most natural resources. During the 1990s the pricing of resources was reformed, and the prices were adjusted to international levels. This is still the case today. In particular, the domestic oil price is adjusted based on the oil markets in Singapore, Rotterdam, and New York, and fluctuates with market demand. The domestic natural gas price is lower than the international price, but it is still determined by the market.

The quality of institutions in China varies significantly over provinces. Chinese provinces possess homogeneous constitutional and legal systems, but their institutional quality differs widely for historical, regional, political, and other reasons. For example, coastal provinces typically have better institutions than inland provinces, partly because they are more open, and partly because some coastal provinces enjoy preferential treatment since the 'reform and open policy' initiated in 1978. These special features help us to study the interplay of resource abundance, institutional quality, and economic growth.

We try to answer two questions. The first question is whether and how the effect of resource abundance on economic growth depends on institutional quality. It is, of course, possible that the resource effect on economic growth may also be dependent on some other variables besides institutional quality, such as manufacturing, R&D, education, and so on (see Fan et al., 2012). We shall focus on the interaction effect of institutional quality, and try to provide explanations how and why institutional quality influences the resource effect on economic growth. The second question is whether the association between resources and economic growth varies over time. Since the West China Development Drive had an emphasis on natural resources, the resources in the Western provinces were exploited more intensively. The economic growth rate in the Western provinces has indeed increased after the Drive was initiated. It is therefore possible that the association between resources and growth is different before and after the policy change, a hypothesis that will be formally tested.

4.3. Data and descriptive statistics

We consider 28 mainland 'provinces' in China, namely 22 provinces, 4 municipalities directly under the central government, and 2 autonomous regions. Three autonomous regions (Xinjiang, Inner Mongolia, and Tibet) are excluded because of lack of data. Each province is labeled either 'West' or 'East' depending on its geographic location; see Figure 4.1.

In studying natural resource abundance and its effect on growth, we distinguish between a stock measure (RA_s, resource reserves) and a flow measure (RA_f, resource revenues). Resource reserves are a measure of *in situ* resource wealth, while resource revenues measure the flow of income from extracting resource stocks at some point in time. Although the two measurements are likely to be highly correlated, the distinction is useful because some provinces may be rich in resource reserves, while their income does not depend primarily on resource exploitation. Also, it is not clear whether resources in the ground have the same effect on economic growth as flows of resource revenues do (Norman, 2009). Both measures differ from resource dependence, the typically used

Figure 4.1: Map of China

Note: The provinces left of the black solid line are defined as Western regions, most of which are affected by the West China Development Drive policy. More precisely, the West China Development Drive policy covers Chongqing, Sichuan, Guizhou, Yunan, Tibet, Shaanxi, Gansu, Qinghai, Ningxia, Xinjiang, Inner Mongolia, and Guangxi. Source: Chinasource website, http://www.chsource.org/site/index.php. Although Shanxi province is usually treated as a central region, it is grouped here as a Western region since its economic structure is more like the Western provinces and it is covered by the West China Development Drive policy. Guangxi province is defined here as a Western province for the same reason.

(though not very good) proxy for resource abundance.

The question whether resource abundance is exogenous or endogenous has been emphasized by Brunnschweiler and Bulte (2008). Van der Ploeg and Poelhekke (2010) pointed out that abundance may not be as exogenous as it seems, and suggested that the historic resource stocks used by Norman (2009) are less endogenous than other measures. We agree with this suggestion and follow Norman (2009) in measuring resource reserves as the recent (2003) observed level of reserves plus total production during the preceding years, including both energy and mineral resources. The energy resources include petroleum, natural gas, and coal mining. The mineral resources cover all major mineral resources in China and include iron ore, manganese ore, chrome ore, vanadium ore, native ilmenite, copper ore, lead ore, zinc ore, bauxite, magnesite, pyrite, phosphate ore, and kaolin. All resource data at the regional level are taken from the China Statistical Yearbook. Because of lack of data in the early years, we can only construct the stock values in 1999 using 1999 prices of resources. Stock values rather than physical quantities are used to enhance comparability across resources, as suggested by Norman (2009). Although stock values may vary a little depending on the price, exploitation technology, and other factors, values in the early years are preferred because they are likely to influence government behavior in later years (Norman, 2009; Van der Ploeg and Poelhekke, 2010). Measuring resource reserves in this way should mitigate (but not eliminate) the endogeneity.

Resource revenues are measured by sales income of resources (after adjusting for inflation), and also include energy and mineral resources. Besides the resources covered in the measure of resource reserves, resource revenues include additional resources such as subterranean heat (energy), nickel ore, tungsten ore, and tin ore (mineral). These additional resources account for less than 20% of sales income on average. (The sales income documented in the Yearbook is the sum of all types of resources, and there are no statistics that allow us to compute sales income covering exactly the same types of energy and mineral resources as the reserves measure.) Since reserves and revenues include almost the same types of resources, we largely rule out the possibility that their difference lies in the different types of resources that they measure.

Compared to other measures used in the literature, our resource abundance measures are less affected by other economic activities, and thus serve as better proxies of resource abundance. We are, however, aware of some weaknesses of our measures. For example, even resource reserves tend to be measured as economically recoverable reserves and are thus subject to changes in prices and technology. Besides, we cannot recover the resource stocks in some of the early years, e.g. 1990, due to lack of data. If these data were available, this would reduce the endogeneity of the resource reserves measure. As for resource revenues, one worry is whether our results will be affected by considering production cost. We check this by experimenting with different measures of revenues, in particular net profit of resources and gross industrial output of resources. Estimation results based on different measures are highly consistent. These measures are also highly correlated (correlation > 0.94), suggesting that production costs differ only marginally across provinces. Therefore we will present our results using sales income of resources, because it is the most complete measure without missing values. The resource revenues measure may be less exogenous than reserves due to market conditions, but the two measures are closely related since resource stocks can be converted into flows of money (Brunnschweiler and Bulte, 2008). The time-varying feature of the revenues measure allows us to examine the short-run (dynamic) relation between resources and growth, while the reserves measure is time-invariant.

The economy-related variables include economic growth, institutional quality, R&D, industrial development, private sector employment, foreign investment, and initial economic level. All of these, except institutional quality and initial economic level, contain observations over several years for each province. The time span varies across variables.

We emphasize the role and measurement of institutional quality. This is a difficult concept to measure. In cross-country studies one typically uses systematic indicators such as the rule of law or government competitiveness (Knack and Keefer, 1995). But within China the constitution and law in one province is the same as in another province. In Chinese studies, institutional quality is therefore often ignored or measured using dubious proxies, such as the ratio of total trade over GDP (Xu and Wang, 2006; Shao and Qi, 2009). We propose to use the confidence in courts surveyed in 1995 by the World Bank as a measure of institutional quality. This is a subjective measure, reflecting perceptions of people from 114 cities. Chinese courts are divided into four levels. The highest level is the supreme court in Beijing, and the other three levels are the so-called people's courts: high courts, intermediate courts, and basic courts. Appointments at the different levels of the people's courts are made by corresponding strata of the people's congresses. Therefore, unlike most of western countries where courts and government have independent power, local courts in China are often influenced by local power cliques. The confidence in courts therefore reflects not only the perceived justice of the courts but also the behavior of the government, and thus captures the essence of institutional quality. The subjectivity of the proposed measure is a potential weakness in that it sometimes differs from an objective measure and could be biased, as suggested by Olken (2009) in a different context. Such a difference or bias (if present) would however be largely averaged out, since we work with aggregated provincial data. An advantage of the subjective measure is that it is based on the perception of several aspects of government behavior, and thus reflects many aspects of institution. It is therefore more general than a specific objective measure, which typically captures only one aspect of government behavior, e.g. corruption, efficiency, or intervention in the economy. Our measure is also likely to be more stable, because it is formed over a period of time, and thus reflects underlying features of local institutions that are not easily changed in the short run. This is especially relevant in rural China, where people are not well-informed about the latest changes of government behavior, and therefore do not rapidly adjust their perceptions, once formed.

The measurement of all variables and their time span is briefly described below.

G Growth of real GDP per capita. In the cross-section analysis, growth is averaged between 1990 and 2008:

$$G = \frac{\log(\mathrm{GDP}_T/\mathrm{GDP}_{T_0})}{T - T_0},$$

where $T = 2008$ and $T_0 = 1990$. In the panel-data analysis, it is defined as the annual growth rate of real GDP per capita:

$$G_t = \log(\mathrm{GDP}_t/\mathrm{GDP}_{t-1}) \qquad (t = 1991, \ldots, 2008).$$

RA$_s$ Log of resource reserves in Chinese Yuan per capita. The variable is constructed by first summing the per capita stock values of all types of resources, and then taking the logarithm. The stock value of a resource is the product of its reserves and its average market price in the corresponding year. The reserve values are constructed using an estimate of the reserves in 1999, obtained by adding extraction flows from 1999 to 2003 to the 2003 'reserve base'. The resource reserves cover energy resources and mineral resources (types of energy and mineral resources are given above).

RA_f Log of resource revenues in Chinese Yuan per capita. The resource revenues are measured by sales income of resources after adjusting for inflation, covering both energy and mineral resources, from 1999 to 2008.

INS Institutional quality, measured by confidence in the courts, which is a weighted average of city level data. The weights are given by the proportion of the city's GDP in the province. (We also used the proportion of a city's population as weights as a robustness check.) Only cross-section data in 1995 are available.

R&D Research and development, the ratio of government expenditure in R&D to total government expenditure, from 1995 to 2006.

IND Industrial development, ratio of value-added of industry to GDP, from 1992 to 2008.

PSE Private sector employment, also referred to as private economic activity, measured by the number of people in not-state-owned companies divided by the provincial population, from 1992 to 2008.

FI Foreign investment proportion, the ratio of the actual inflow of foreign investment over gross investment in fixed assets, from 1989 to 2003. This captures the importance of foreign investment in the local economy.

INIT Initial economic level in Chinese Yuan per capita, defined as the logarithm of real GDP per capita in 1989.

WEST Geographic dummy: WEST = 1 if the province lies in the Western region of China (as defined in Figure 4.1) and WEST = 0 otherwise.

The resource reserves data are taken from the China Statistical Yearbook (National Bureau of Statistics, 2003–2004) and the China Economic Information Network (CEINET). Resource revenues data are taken from the China Land and Resources Statistical Yearbook (Ministry of Land and Resources, 2000–2009). The economy-related variables are either from CEINET or from the World Bank.

Table 4.1 provides descriptive statistics of the economy-related and resource variables. By comparing the statistics of the East sample to the West sample, we see that the average growth rate in Eastern provinces is generally higher than in Western provinces. On the other hand, Western provinces have slightly higher resource reserves and revenues

Table 4.1: Descriptive statistics of economy-related variables

	Entire sample (28 prov)		East sample (18 prov)		West sample (10 prov)	
Variable	Mean	Std	Mean	Std	Mean	Std
Growth	0.0394	0.0054	0.0411	0.0054	0.0364	0.0039
RA_s	4.4511	1.0375	4.3193	1.2015	4.8292	0.6398
RA_f	2.0758	0.3855	2.0238	0.4224	2.2522	0.3584
INS	58.146	13.321	62.009	11.758	51.199	12.963
R&D	0.0097	0.0046	0.0111	0.0053	0.0073	0.0013
IND	0.2875	0.0665	0.2973	0.0752	0.2698	0.0452
PSE	0.0655	0.0278	0.0763	0.0283	0.0461	0.0126
FI	0.0963	0.0928	0.1305	0.0991	0.0347	0.0266
INIT	3.1876	0.2068	3.2626	0.2118	3.0527	0.1100

than Eastern provinces. The institutional quality is generally better in the Eastern provinces than in the Western provinces. Also, R&D, industrialization, private sector employment, and foreign investment in the East are all higher on average than in the West.

4.4. Cross-section analysis

We analyze the data first as a cross section, and then, in Section 4.5, as a panel. We begin by reconsidering the classical growth regression

$$G = \beta_0 + \beta_1 \mathrm{RA} + \beta_2 \mathrm{INS} + \sum_{k=1}^{6} \theta_k x_k + \epsilon_1, \tag{4.1}$$

where G denotes economic growth, RA is resource abundance, INS represents institutional quality, and

$$\big(x_1, \ldots, x_6\big) = \big(\mathrm{R\&D}, \mathrm{IND}, \mathrm{PSE}, \mathrm{FI}, \mathrm{INIT}, \mathrm{WEST}\big)$$

contain auxiliary control variables: research and development (R&D), industrial development (IND), private sector employment (PSE), foreign investment (FI), initial economy level (INIT), and the Western dummy (WEST). The auxiliary variables affect the economy and are associated with resource abundance. Their inclusion will therefore reduce the omitted variable bias. For resource abundance, we always consider two variants: one where RA is measured as a stock (RA_s, resource reserves) and one where it is measured as a flow (RA_f, resource revenues).

Table 4.2: Economic growth: classical growth model

	(a)	(b)	(c)	(d)	(e)	(f)
RA_s	-0.0001		0.0002		-0.0037	
	(-0.22)		(0.26)		(-0.76)	
RA_f		0.0001		0.0006		-0.0099
		(0.08)		(0.26)		(-1.36)
INS	0.0002	0.0002	0.0001	0.0001	-0.0002	-0.0003
	(2.76)	(2.76)	(1.85)	(2.02)	(-0.52)	(-0.96)
R&D			0.3074	0.3011	0.3921	0.4391
			(1.87)	(2.03)	(1.74)	(1.96)
IND			0.0217	0.0210	0.0172	0.0194
			(1.96)	(1.92)	(1.32)	(1.76)
PSE			0.1359	0.1320	0.1583	0.1417
			(2.04)	(2.43)	(2.23)	(2.62)
FI			0.0258	0.0269	0.0284	0.0279
			(2.76)	(2.56)	(2.94)	(2.65)
INIT	-0.0053	-0.0051	-0.0304	-0.0296	-0.0345	-0.0328
	(-1.06)	(-0.99)	(-4.07)	(-4.68)	(-3.87)	(-5.39)
WEST	-0.0035	-0.0036	-0.0016	-0.0016	-0.0014	-0.0012
	(-1.29)	(-1.29)	(-0.62)	(-0.58)	(-0.52)	(-0.45)
RA $\times$ INS					0.0001	0.0002
					(0.82)	(1.39)
Constant	0.0465	0.0448	0.1089	0.1066	0.1392	0.1392
	(2.38)	(2.12)	(6.18)	(5.47)	(3.28)	(4.79)
R^2	0.4579	0.4574	0.6829	0.6832	0.6913	0.6996
p-value of F-test	0.0022	0.0022	0.0000	0.0000	0.0000	0.0000

Note: t-values are in parentheses. The number of observations in each column is 28.

The least-squares estimation results are presented in Table 4.2. Columns (a) and (b) show that the impact of resource abundance is very weak, both when measured as a stock (resource reserves) and when measured as a flow (resource revenues). The effect of institutional quality is strong and positive. If other explanatory variables are added (columns (c) and (d)), then the significance of institutional quality slightly decreases but it remains strong, while the resource effect remains insignificant. Equation (4.1) seems to imply that resource abundance has no effect on economic growth, but such a conclusion would be premature and incorrect. Classical growth regressions cannot fully capture the resource effect in China, because the resource effect is likely to vary with institutional quality. It is possible that natural resources are important in provinces with poor institutional quality, but less important in provinces with strong institutional quality. Classical growth regressions ignore such provincial heterogeneity by assuming constant coefficients for each explanatory variable. The estimated coefficient in the presented classical growth regression is the 'overall' effect of resource abundance, and its insignificance does not imply that heterogenous effects are also insignificant for various levels of institutional quality. In fact, we shall see that provincial heterogeneity is essential in explaining the role of resource abundance.

Thus motivated, we extend the classical models by including an interaction term RA×INS, as suggested by Mehlum et al. (2006). We estimate the regression model

$$G = \beta_0 + \beta_1\mathrm{RA} + \beta_2\mathrm{INS} + \beta_3\mathrm{RA}\times\mathrm{INS} + \sum_{k=1}^{6}\theta_k x_k + \epsilon_2. \tag{4.2}$$

The estimation results are given in columns (e) and (f) of Table 4.2. We see that the interaction term is positive, but not significant. This result weakly supports the argument by Mehlum et al. (2006) that resource abundance promotes the economy if institutions are producer-friendly. The insignificance of the interaction term suggests that the linear model may not fully capture the interaction effect of resources and institutional quality in China. Note that Equation (4.2) only provides a positive or negative (linear) interaction effect, and that this effect is the same for all institutional quality levels. However, if resource effects on growth depend *nonlinearly* on institutional quality, then (4.2) does not capture this.

In order to capture a possibly nonlinear relationship between resource abundance and economic growth, we consider the functional coefficient model

$$G = \delta_0 + \delta_1\mathrm{RA} + \sum_{k=1}^{6}\gamma_k x_k + \epsilon_3. \tag{4.3}$$

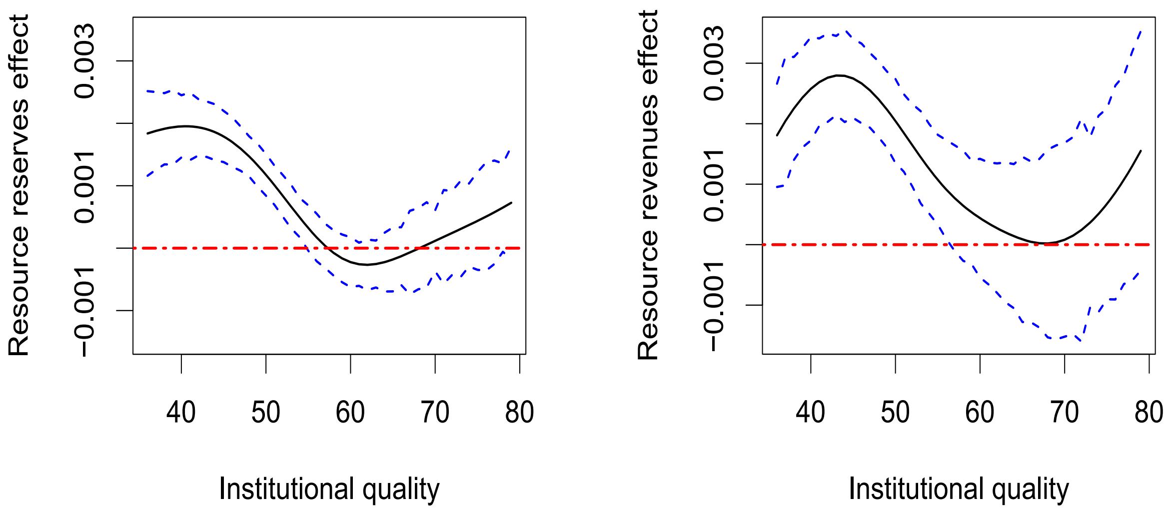
The same variables RA and $x_1, \ldots, x_6$ appear in Equation (4.3) as in Equation (4.1), except that institutional quality INS enters through the coefficients δ_0, δ_1, and γ_k ($k = 1, 2, \ldots, 6$). Since there is no *a priori* reason why some of the coefficients would and others would not depend on institutional quality, we allow all coefficients to be functions of institutional quality. The advantage of a functional-coefficient model is that it provides information on how the interaction varies (possibly nonlinearly) across different levels of institutional quality. A second advantage is that it solves the potential reverse causality between institutional quality and growth, at least to some extent, because institutional quality enters the model as a smoothing variable instead of a control variable.

The parameters in this model are estimated by local linear estimation (Fan and Gijbels, 1996; see also Cai et al., 2000). Thus, we specify

$$\begin{aligned} \delta_j &= \delta_{Cj} + \delta_{Sj}(\mathrm{INS} - u_0) \quad (j = 0, 1), \\ \gamma_k &= \gamma_{Ck} + \gamma_{Sk}(\mathrm{INS} - u_0) \quad (k = 1, 2, \ldots, 6), \end{aligned}$$

where min(INS) $\leq u_0 \leq$ max(INS). The parameters $(\delta_{Ck}, \delta_{Sk})$ and $(\gamma_{Ck}, \gamma_{Sk})$ are estimated nonparametrically. Various data-driven methods can be employed for selecting the bandwidth. We chose the bandwidth by minimizing the average mean squared error (Cai et al., 2000).

Figure 4.2: Marginal effect of RA_s and RA_f on economic growth as a function of institutional quality



In Figure 4.2 we show how the δ_1-parameter changes as a function of institutional quality. The solid line plots the estimate of δ_1, and the dashed lines are 5% confidence intervals based on jackknife standard errors. We see that resource reserves and resource revenues largely measure the same concept of abundance (when applied to growth regressions in China). The typical 'U-shape' in both subfigures shows strong and positive correlation between resource abundance and economic growth in provinces with weak institutional quality. As institutional quality improves, this correlation decreases and becomes statistically insignificant. These results provide an explanation of the insignificance of the resource effect in Equation (4.1). The reason for the insignificant 'overall' effect (columns (a)–(d) in Table 4.2) is that the resource effect varies with institutional quality and that this effect is weak in provinces with good institutional quality. The nonlinear behavior in both subfigures also explains the statistical insignificance of the interaction term (columns (e) and (f) in Table 4.2).

In general, resource abundance in China thus has a *positive* effect on economic growth. This evidence obviously challenges the existence of a resource curse. In fact it supports

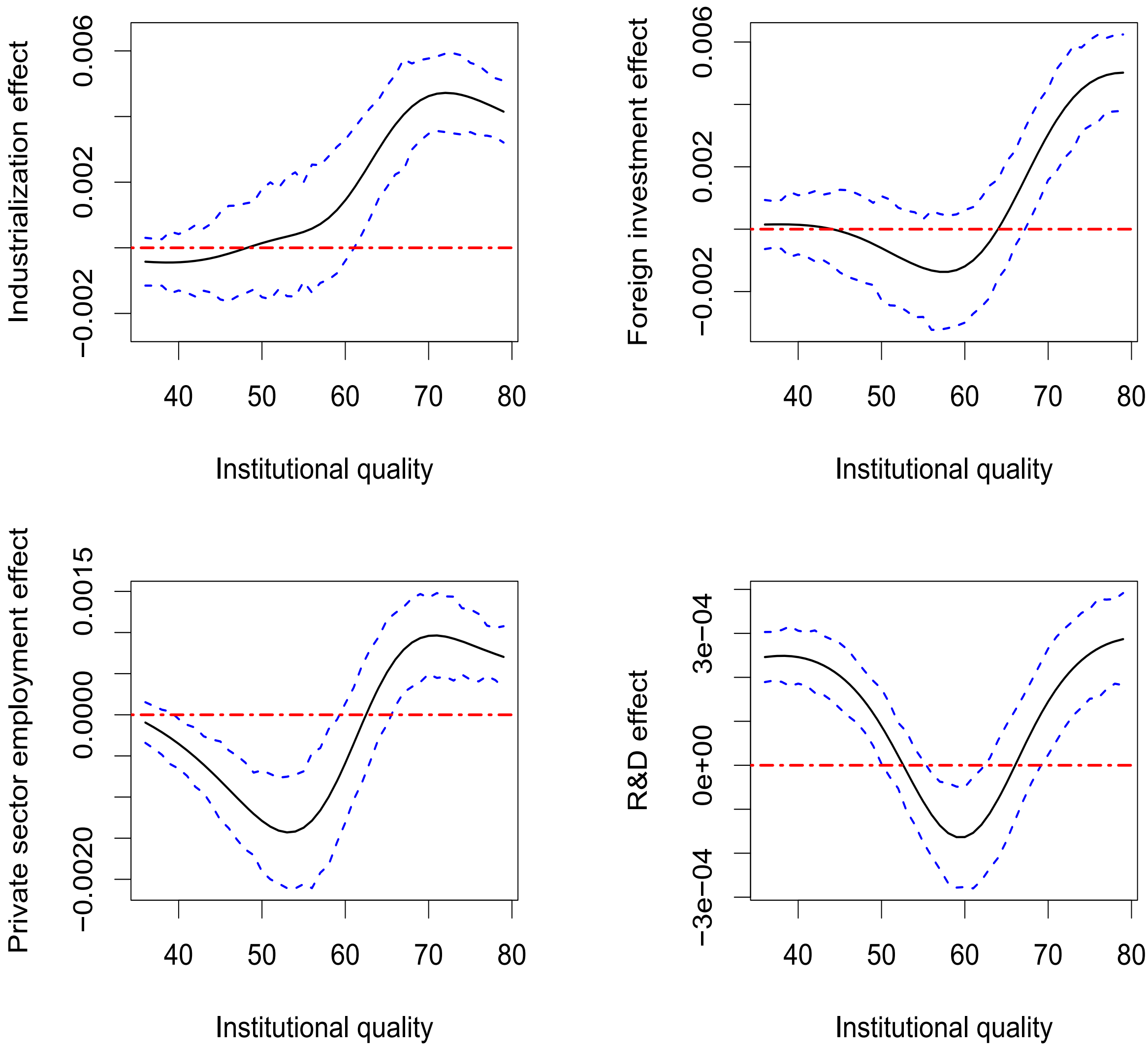
Brunnschweiler and Bulte's (2008) argument that resource abundance promotes economic growth, which they explain by the 'windfall' flow of income from resource extraction. This flow, they argue, has a direct effect on the economy as well as an indirect effect through improving institutional quality.

The positive effect of resource abundance is particularly strong in regions with weak institutional quality, and the effect decreases as institutional quality improves. This finding differs from the cross-country evidence reported by Mehlum et al. (2006), who find that worse institutions make the effect of natural resources more negative. A possible explanation is that regions with weak institutional quality are likely to rely more on their primary industries than regions with strong institutional quality, because the prosperity of many non-resource sectors is largely built on good institutional quality. For example, good institutions lead to more willingness of savers to invest in firms and to a higher effectiveness of corporate governance, thus associating good institutions with a healthy financial sector (Beck and Levine, 2005). Nunn (2007) pointed out that better contract enforcement makes countries more specialized in the industries in which so-called relationship-specific investments play a dominant role.

Improvement of institutional quality helps the development of many non-resource sectors more than it helps the development of resource sectors, so that better institutions make resources become less important. The correlation between the economy and non-resource sectors is thus stronger than correlation between the economy and resource abundance, and indeed we observe a decreasing and insignificant effect of resource abundance when institutional quality increases.

To further strengthen this argument, we present the functional effects of some other control variables, namely industrialization, foreign investment, private sector employment, and R&D. Figure 4.3 shows the effects of industrialization, private sector employment, and foreign investment on economic growth all tend to be stronger and more positive in regions with better institutional quality. Typical examples are Qinghai, Guizhou, and Ningxia. These provinces all suffer from weak institutional quality, and their economies therefore rely largely on resource abundance, while the non-resource sectors are poorly developed. In contrast, Guangdong, Jiangsu, Zhejiang, and Tianjing provinces are among the top ten provinces in terms of institutional quality, and their non-resource sectors, such as R&D, industrialization, and private sectors are among the

Figure 4.3: Marginal effect of other control variables on economic growth as a function of institutional quality



best. Resources in these provinces play only a small role in promoting economic growth.

When institutional quality exceeds the median level (62 on the left, 68 on the right in Figure 4.2), the positive impact of resource abundance on economic growth increases (but remains insignificant) as institutional quality improves. Apparently, provinces with strong institutional quality *and* abundant natural resources are likely to make good use of these resources and revenues. Property rights on natural resources in China are owned by the government, and local residents therefore typically do not benefit much from revenues derived from resource extraction. Most income associated with resources goes to the government and to state-owned enterprizes. Hence, for provinces with weak institutional quality, rising revenues from the booming resource sectors are not used by the government to stimulate the economy, but often harm economic development, because they lead to increased prices for nontradable goods, thus lowering the competitiveness of local economies (Zhang et al., 2008). If institutional quality is strong, however, then revenues from resources may be used to boost economic development. This is because better property rights tend to improve asset allocation, leading to higher growth (Claessens and Laeven, 2003). Examples include Shandong, Jilin, Liaoning, Tianjin, and Henan provinces, most of which are traditional industrial provinces in North-East China. These provinces are rich in natural resources (especially mineral resources), the institutional quality is high, and the exploitation and use of the resources is efficient. Booms in resource sectors thus do not impede the development of non-resource sectors, but instead stimulate industries that are indirectly related to resources such as the automobile, shipbuilding, and equipment manufacturing industries.

Our results are very robust to different measures of resource revenues and to different weights for institutional quality. Even when we take into account the possible influence of the 2000 policy shock and estimate the model with separate samples (before and after the shock), the results remain essentially unchanged. (Further details on the shock and its effects will be discussed in the next section.) We conclude that the effect of resources on the economy is highly and nonlinearly dependent on institutional quality, and that the correlation between resource abundance and economic growth is high and positive in provinces with weak institutional quality, but weakly negative in provinces with medium institutional quality.

4.5. A panel-data approach with time-varying resource effects

Cross-section models are useful in capturing long-run effects, but they do not identify resource effect fluctuations over time. In particular, they cannot identify the effects of a policy shock such as the West China Development Drive. This significant policy package was introduced by the Chinese central government in 2000 with the purpose of stimulating the economy of the Western regions. While the policy also involves non-resource projects (such as promoting infrastructure construction, protecting the ecology environment, and re-adjusting industrial structure), the emphasis is on intensifying natural resource exploitation. Several significant projects connected with natural resources in West-China have resulted from this initiative. For example, the West-East natural gas transmission project led to an increase of natural gas production in Sichuan and Qinghai provinces by more than 100% and 900%, respectively, between 2000 and 2007. Also, steel production in Yunan and Guizhou provinces increased by around 200% and 400%, respectively, since the Drive began. The economic growth rate in Western provinces has indeed increased since 2000. It is therefore to be expected that the relationship between resources and growth is different before and after the policy, and we shall test this hypothesis using both a standard and a time-varying coefficient panel-data approach. The availability of panel data is important because, by allowing not only variation over provinces but also over time, we are able to capture the short-run dynamic behavior of the resource effect as well as other impacts on economic growth. In addition, the use of panel data enlarges the sample size and hence improves the accuracy of the estimates. Since the stock measure of resource abundance is a historical variable that does not vary over time, we can only use the flow measure (resource reserves) in the panel-data analysis.

4.5.1. Standard panel-data approach

We first consider the standard panel-data model. To incorporate the policy shock in 2000, we introduce a policy shock dummy PD taking the value 0 before 2000 and 1 from 2000 onwards. Thus we have

$$G_{it} = c_i + \beta_0 + \beta_1 \mathrm{RA}_{f\,it} + \beta_2 \mathrm{PD}_{it} + \beta_3 \mathrm{RA}_{f\,it} \times \mathrm{PD}_{it} + \sum_{k=1}^{4} \theta_k z_{k,it} + \epsilon_{it},$$

where $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and we allow for the possibility of an interaction term $\mathrm{RA}_f \times \mathrm{PD}$. Here G_{it} denotes the growth rate of real GDP per capita in province i

at year t, c_i is a province-specific effect, and PD and $\text{RA}_f \times \text{PD}$ capture the policy effect. The auxiliary control variables in this case are $(z_1, \ldots, z_4) = (\text{R\&D}, \text{IND}, \text{PSE}, \text{FI})$. The idiosyncratic error ϵ_{it} is assumed to be independent of x_{it}. Since province-specific effects are correlated with the regressors, we employ a fixed-effect estimation method. The time-invariant variables INS, INI, and WEST are excluded as explanatory variables, because they cannot be identified in a fixed-effect method. Since our measure of institutional quality varies only slightly in our observed time period (see Section 4.3), the exclusion of INS will only have a slight effect on the results. We only use resource revenues (the flow) as a measure of resource abundance, because our measure of resource reserves (the stock) does not vary over time.

Table 4.3: Economic growth: standard panel-data model

	(a)	(b)	(c)	(d)	(e)
RA_f	0.0187	0.0132	0.0072	0.0128	0.0107
	(5.51)	(4.16)	(2.16)	(4.38)	(4.94)
R&D	−0.9825	−0.9486	−0.8366	−0.8938	−0.9166
	(−3.80)	(−4.83)	(−3.42)	(−3.72)	(−3.91)
IND	0.0854	0.0734	0.0635	0.1039	0.1063
	(5.66)	(5.00)	(3.86)	(8.14)	(8.18)
PSE	0.0283	0.0170	0.0305	0.0179	0.0293
	(0.57)	(0.33)	(0.81)	(0.35)	(0.67)
FI	−0.0510	−0.0216	−0.0214	−0.0282	−0.0276
	(−1.84)	(−0.92)	(−1.07)	(−1.17)	(−1.17)
PD		0.0105	−0.0059		
		(5.08)	(−0.57)		
$\text{RA}_f \times \text{PD}$			0.0091		
			(1.67)		
D_{0304}				0.0154	−0.0106
				(3.55)	(−0.48)
$\text{RA}_f \times \text{D}_{0304}$					0.0124
					(1.10)
Constant	−0.0050	−0.0004	0.0108	−0.0034	−0.0064
	(−0.66)	(−0.07)	(1.40)	(−0.54)	(−0.11)
overall R^2	0.1534	0.2163	0.2279	0.2468	0.2547
p-value of F-test	0.0000	0.0000	0.0000	0.0000	0.0000
ρ	0.2772	0.2026	0.1948	0.2349	0.2432

Note: t-values are in parentheses. ρ is the fraction of variance due to the individual-specific effect. The number of observations in each column is 308.

Table 4.3 presents the standard fixed-effect estimation results. Column (a) is the benchmark, including only resource revenues and auxiliary control variables. The association between resource revenues and growth is positive and strong in the short-run dynamics, contrasting sharply with the insignificant long-run relationship. This discrepancy between short-run and long-run effect is in line with Collier and Goderis (2008) who find similar differences in a cross-country framework. One explanation of this positive association is the income effect that the 'windfall gain' from resources stimulates consumption and further prompts economic growth. As expected, R&D and foreign investment have different short-run and long-run effects; these are long-run investments having little effect in the short run. To study whether the 2000 policy shock affects economic growth, column (b) includes the policy shock dummy PD. Its coefficient is strongly significant, showing, as expected, that the new policy has led to higher growth. If we include both PD and the interaction term $\mathrm{RA}_f \times \mathrm{PD}$, then we obtain column (c). The interaction term is positive and weakly significant (p-value is 9.5%), suggesting the possibility that the resource effect is different before and after the policy shock. The result is not conclusive however, because a standard panel-data model can only measure the linear difference between before and after the shock, thus capturing the average change. If the resource effect contains nonlinear dynamics, then these are not captured by the standard panel-data model. This leads us to the time-varying coefficient model, where possibly nonlinear resource effects can be investigated. This model and the resulting columns (d) and (e) are discussed in the next subsection.

4.5.2. Time-varying coefficient approach

The standard fixed-effect approach reveals different resource effects before and after the policy shock. This approach can not, however, describe how the resource effect changes after the policy shock. We expect that the effects of other variables are also influenced by the policy. This could lead to a strengthening of the effects, because the policy also involves non-resource projects and these non-resource industries may grow faster after 2000. But it could also lead to a weakening, because the emphasis on resource exploitation strengthens the association between resources and economic growth, and also because an over-emphasis on resources crowds out the non-resource sectors, leading to a weaker correlation between non-resource sectors and growth.

We extend the standard panel-data model by allowing the coefficients to be time-varying, and consider the time-varying coefficient model

$$G_{it} = c(t) + x_{it}'\tau(t) + \epsilon_{it} \qquad (i = 1, \ldots, N, \;\; t = 1, \ldots, T),$$

where x_{it} are the explanatory variables: RA_f, R&D, IND, PSE, and FI, all in province i at year t. The dummy PD is excluded because the policy effect can now be captured by the model parameters which are smooth functions of t. The parameter vector $\tau(t) = \{\tau_1(t), \tau_2(t), \ldots, \tau_k(t)\}'$ contains the coefficients for the $k = 5$ control variables. This is a typical time-varying coefficient model for panel data (Hoover et al., 1998). Unlike the standard panel-data model, no within-transformation or first-difference transformation is needed when estimating the model, because no incidental parameter problem occurs in this case.

To estimate $c(t)$ and $\tau(t)$, we employ a method proposed by Hoover et al. (1998, Section 2.3), in particular the local constant fit based on a kernel function $K(\cdot)$ with bandwidth h. The kernel estimator of $c(t)$ and $\tau(t)$ is then given by

$$\begin{pmatrix} \hat{c}(t) \\ \hat{\tau}(t) \end{pmatrix} = \left(\sum_{i=1}^{N} X_i' K^*(t) X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' K^*(t) G_i \right),$$

where

$$X_i = \begin{pmatrix} 1 & x_{i1,1} & \ldots & x_{i1,k} \\ 1 & x_{i2,1} & \ldots & x_{i2,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{iT,1} & \ldots & x_{iT,k} \end{pmatrix}, \qquad G_i = \begin{pmatrix} G_{i1} \\ G_{i2} \\ \vdots \\ G_{iT} \end{pmatrix}$$

are a $T \times (k+1)$ matrix and a $T \times 1$ vector, respectively, and

$$K^*(t) = \begin{pmatrix} K((t-1)/h) & 0 & \ldots & 0 \\ 0 & K((t-2)/h) & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & K((t-T)/h) \end{pmatrix}$$

is a diagonal $T \times T$ weight matrix. The bandwidth is selected following Hoover et al. (1998, Section 2.4) by minimizing the average predictive squared error with 'leave-one-out' cross-validation.

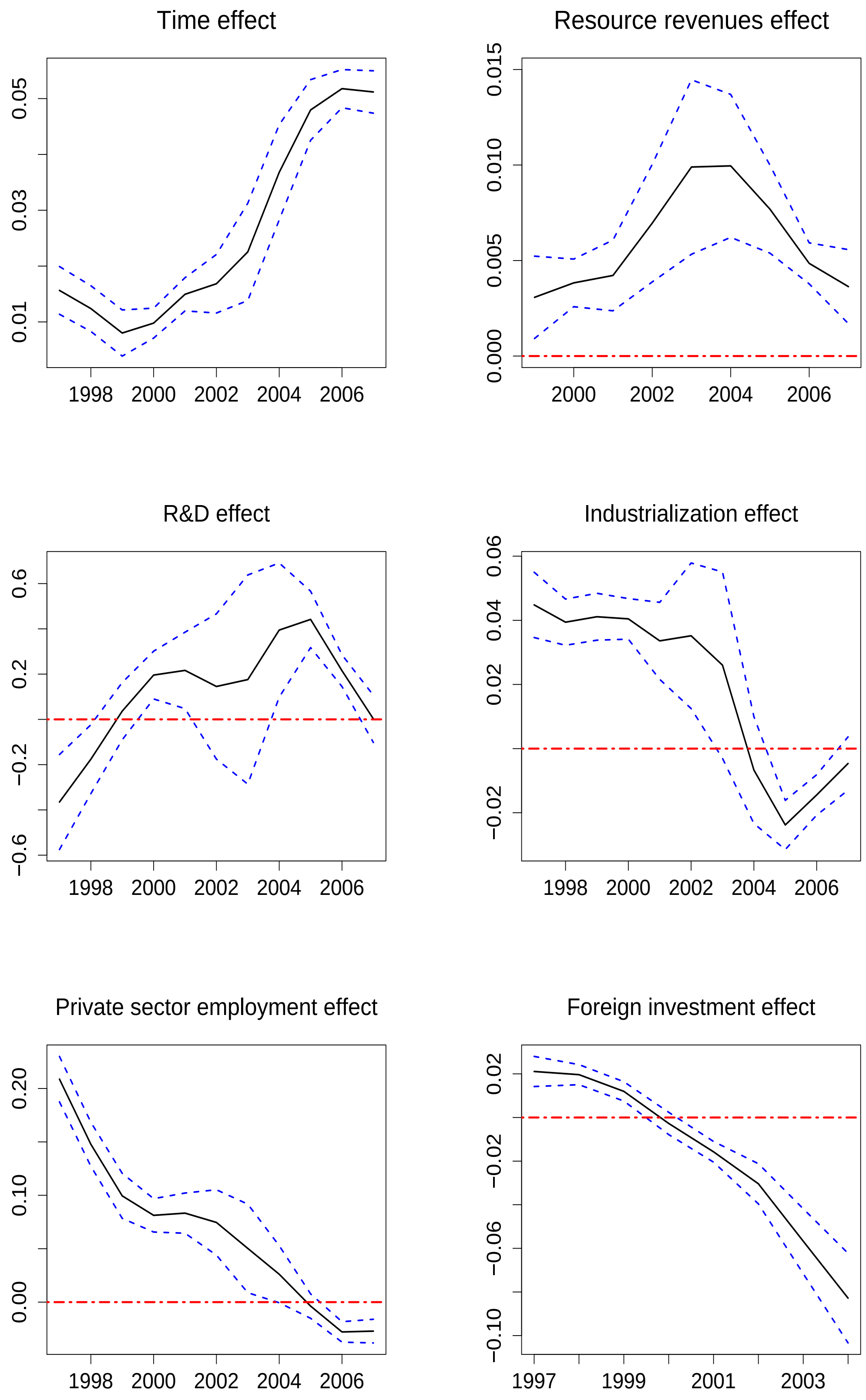
The kernel estimator $(\hat{c}(t), \hat{\tau}(t))$ thus takes the form of a generalized least-squares estimator with weight matrix $K^*(t)$. Rather than running a cross section for every time period, the kernel estimator employs not only the information at time t but also the neighboring information, and its smoothness depends on the choice of bandwidth. By selecting an optimal bandwidth, we minimize the average predictive squared error, and obtain estimators with appropriate smoothness. The kernel estimator has also attractive asymptotic properties (Hoover et al., 1998), but whether these properties apply here is somewhat dubious because of the small number of provinces.

Figure 4.4 shows the results using the entire sample (all provinces). All coefficients vary over time. The coefficient of resource revenues is positive and increasing from 2000–2004, and decreasing from 2005–2007. The R&D effect is increasing from 1997–2000, fluctuating a little during 2001–2005, and decreasing after 2005. The coefficients of industrialization, private sector employment, and foreign investment are generally decreasing from 1997–2004. The estimated time-varying estimates are generally in line with the standard panel-data results (except R&D). In particular, the nonlinearly dynamic resource effect explains the positive but weakly significant coefficient of the interaction term $\mathrm{RA}_f \times \mathrm{PD}$ (column (c) in Table 3). Since the resource effect first increases and then decreases after the shock, the before-and-after difference is partially offset and thus not strongly significant *on average*. But, in general, the resource effect did change after the policy shock, and became considerably stronger immediately after 2000, implying that the correlation between resource revenue and economic growth is stronger after than before 2000. This is not surprising because the emphasis of the West China Development Drive was on exploiting the resources in the Western provinces more intensively and efficiently. Income in these regions has increased, stimulating economic growth, but not equally in all regions. The decreasing coefficients of the other variables suggest that the negative impact of the policy on non-resource effects dominates the positive impact.

In the period 2003–2004 the impact of resource revenues was particularly strong, be it with relatively large standard errors. To confirm this result in the standard fixed-effect model we included a time dummy D_{0304} for 2003–2004, and an interaction term $\mathrm{RA}_f \times \mathrm{D}_{0304}$. Columns (d) and (e) in Table 4.3 show that D_{0304} is significantly positive, confirming that the economic growth rate was particularly high in 2003–2004. The interaction term $\mathrm{RA}_f \times \mathrm{D}_{0304}$ is positive, though not very precise, suggesting that the resource effect increased during the period. In contrast, industrialization, private sector employment, and foreign investment effects experienced a drop in these two years.

Apparently the economic situation in China was different in 2003 and 2004 than in other years, and growth determinants had different effects during this period. This is indeed the case. The economic growth rate was particularly high in 2003 and 2004,

Figure 4.4: Time-varying coefficients: entire sample



Note: The solid curve is the estimated coefficient of each regressor, and the two dashed curves represent ± 1.96 jackknife standard error bands. The jackknife standard errors are computed by leaving out one individual at a time from the sample.

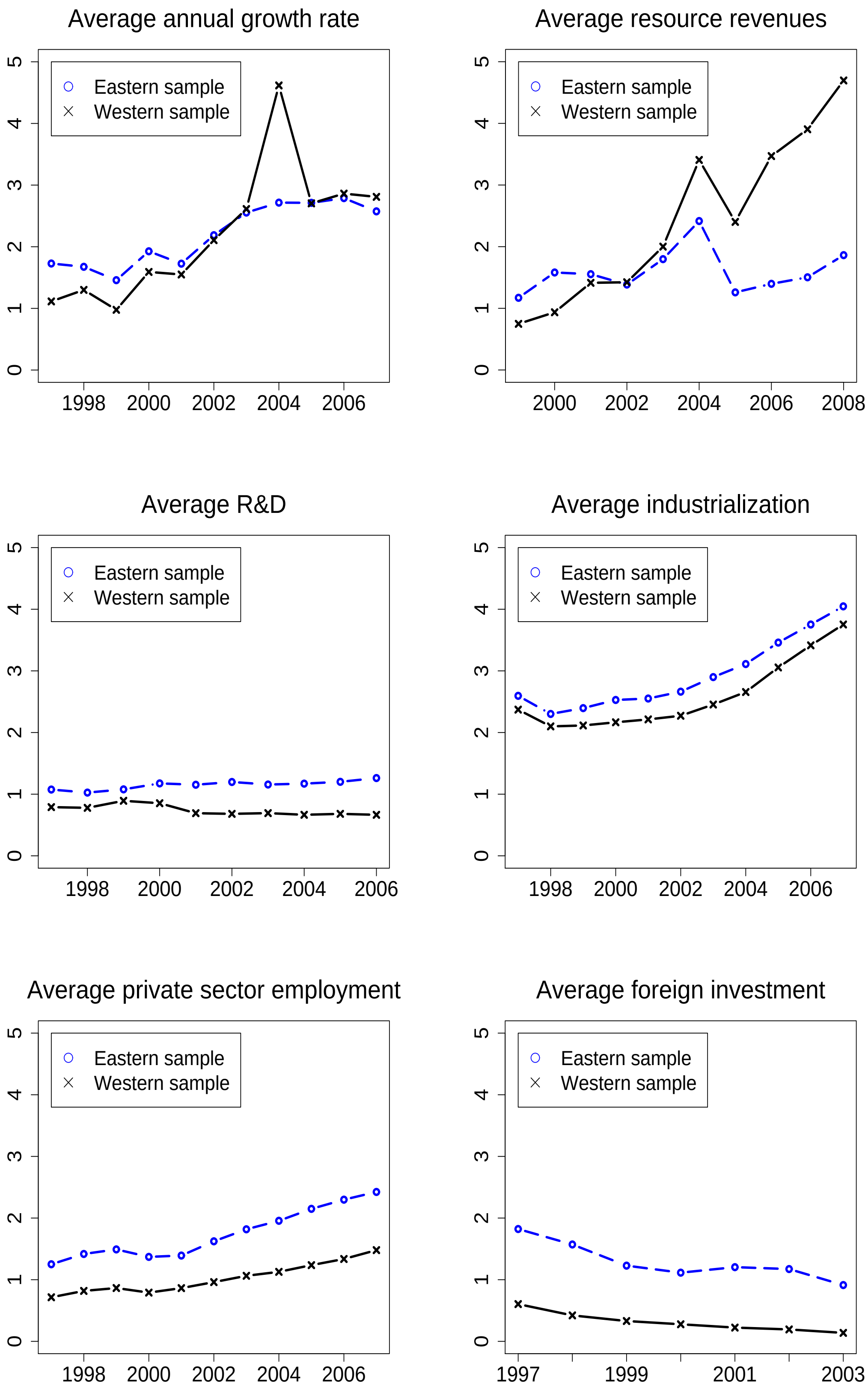
mainly due to a high demand for investment. The annual growth rates of fixed asset investment in 2003 and 2004 were 27.7% and 26.6%, respectively. One reason for such high investments was an increasing demand for housing and automobiles. This demand directly stimulated the investment in the realty business and the automotive industry, and also indirectly in related industries (e.g. steel, building materials, power sector). In addition, several great projects in the Western provinces were initiated in this period: the West-East natural gas transmission project, the West-East electricity transmission project, and the Western coal mining project. These projects led to a large increase in resource production with an associated increase in income. These two reasons explain the strong correlation between resource revenues and economic growth in the period 2003–2004, while the effects of other explanatory variables are relatively weak.

To avoid overcapacity in the future, the government proposed policies to restrain investment. As a result, fixed asset investment largely decreased in 2005 and 2006, and economic growth slowed down. However, resource exploitation did not slow down, and resource revenues kept on increasing, especially in the Western provinces. This is why we observe a decreasing correlation between resource revenues and economic growth rate after 2004.

To understand this from another viewpoint, we plot annual economic growth and its determinants in Figure 4.5. All variables are averaged over Eastern and Western provinces, respectively, and scaled to facilitate comparison. We observe a positive jump in the growth rate in 2004, especially in the Western provinces, and a return to a lower level in 2005. We also observe a jump in resource revenues in both Eastern and Western provinces in 2004. The co-movement of economic growth and resource revenues provides further evidence of the strong correlation between resource revenues and economic growth in 2003–2004. When the economic growth rate returned to a lower level after 2004, resource revenues in the Western provinces were still increasing at a high speed from 2006 to 2008. This explains the decreasing correlation between economic growth and resource revenues after 2004.

The decreasing correlation between economic growth and resource revenues after 2004 shows that increased resource exploitation did not promote the development of other industries and sectors typically regarded as engines of economic growth. As exemplified in Figure 4.5, average R&D, industrialization, private sector employment, and foreign

Figure 4.5: Time series plot of growth and its determinants



investment all changed relatively little as resource revenues increased sharply. Typical examples are Ningxia and Gansu provinces, where resource revenues increased significantly after 2000, but most of the other sectors were still underdeveloped. The economies of the Western provinces still relied much on primitive sectors, and the industrial structure of the Western provinces failed to modernize. The emphasis on resource exploitation brought extra income in the short run, but it did not narrow the gap between West and East China. In addition, only part of the resources produced by the Western regions was used to improve the local economy. The larger part was transported to the Eastern regions to meet the large demand for energy and resources there. For example, the most important gas field in Sichuan province transmitted more than 70% of its natural gas to Eastern provinces. This may also have resulted in enlarging the gap between Eastern and Western provinces. In summary, the intensification of resource exploitation in the Western provinces helped the local economy to some extent, but the positive effect was short-run and not long-run.

4.6. Conclusions

In this paper we have re-examined the effect of natural resource abundance on economic growth at the provincial level in China. We emphasize four features of our analysis. First, we employ new data on natural resource abundance and institutional quality to study the association between resource abundance, institutional quality, and economic growth. We compare two types of resource abundance measures: a stock measure and a flow measure. The new measures of resource abundance are considered to be more exogenous than the conventional resource dependence measure. Institutional quality is measured by a subjective measure of confidence in courts, and it is shown to be theoretically and empirically related to resource abundance and economic growth.

Second, we model a nonlinear resource effect on economic growth. Classical growth regressions cannot fully capture the resource effect on economic growth in China because the resource effect is (nonlinearly) dependent on institutional quality. Thus we employ a functional-coefficient model and we find that the effect of resources on the economy is a nonlinear function of institutional quality, and that the correlation between resource abundance and economic growth is strong and positive in provinces with weak institutional quality, but relatively weak in provinces with strong institutional quality. This finding partially supports the argument in Mehlum et al. (2006) that the resource effect depends on institutional quality, but it suggests that such dependence may not be captured satisfactorily by the linear model considered by them. More importantly, the conclusion that worse institutions make the effect of natural resources more positive (rather than more negative) in China also contrasts with the cross-country evidence in Mehlum et al. (2006).

Third, we study the different roles of resources on economic growth before and after the 2000 policy shock, and find that the association between resources and economic growth is not constant over time if we consider short-run dynamics. Immediately after the 2000 policy shock, the positive correlation between economic growth and resource revenues was increasing, but this did not last long. After 2004 economic growth slowed down while resource revenues kept increasing, leading to weak correlation.

Finally, we analyze the resource effect using both cross-section and panel data. The cross-section model typically captures the long-run effect, and the panel-data model the short-run effect. Abundant resource revenues are positively correlated with economic growth in the short-run, and their long-run correlation is positive in provinces with weak institutional quality.

Although our paper is a cross-province study in China, some ideas can be applied to more general cross-country studies. Our paper suggests that the classical growth model is not always satisfactory in studying resource effects, because it fails to capture a possibly nonlinear influence of institutional quality. It is likely that institutional quality is also relevant in other countries. This is also the case with our finding that the resource effects in China change over time. This is likely to be true in other countries. For example, evidence before World War II tends to support a positive effect of resources on growth (Habakkuk, 1962), while most empirical studies using data after World War II report a negative effect.

Further research is still needed in at least three directions. First, economic growth just measures one aspect of economic development. Economic development also includes *inter alia* a decrease in poverty and infant mortality, and better nutrition (Bulte et al., 2005). In many countries with high growth rates there is poverty and basic nutritional needs are not met. Therefore, the effect of natural resources on economic growth is not necessarily the same as the effect of natural resources on economic development (Zhang et al., 2008). Second, the exogeneity of resource abundance deserves more investigation, and the quality of resource abundance measures may be further improved by using stock values of earlier years. Third, while a more general institutional indicator has been used in cross-country studies, there are no systematic indicators on institutional quality in China. Since institutional quality appears a key variable, more accurate measures would sharpen the analysis and improve the estimates.

Chapter 5

Dress-up contests: A dark side of fiscal decentralisation[4]

5.1. Introduction

During the last three decades, fiscal decentralisation (FD) and local-government reform have been at the centre-stage of policy experiments. This has occurred not only in countries with a traditional tendency to decentralise, such as the United States, but also in a large number of developing and transition economies, such as Africa, Asia, and Latin America (The World Bank, 1999). FD, which moves the responsibility for decision-making in public expenditure from central to local governments, is widely believed to be an effective tool for improving the efficiency of public expenditure. One of the major transmission channels, well documented in the literature, is yardstick competition, through which FD regulates the behaviour of Leviathan government (Besley and Case, 1995; Belleflamme and Hindriks, 2005; Besley and Smart, 2007; Bordignon et al., 2004); see Lockwood (2005) for a recent review.

The literature focuses on the benefits of FD; in contrast we explore a negative aspect. We argue that under asymmetric information, the yardstick competition of *capability* between local governments (due to FD) turns into a competition for a *better image*, i.e. a 'dress-up contest'. This is because voters with limited information cannot observe the politicians' capability; they instead infer this capability from the public service provided. This motivates politicians to allocate more resources to the public goods that can best demonstrate their capability. The dress-up contest can lead to a structural bias in public expenditure, which may result in a distortion of social welfare.

This paper has two main contributions. First, we propose and model dress-up contests that occur between local governments and are caused by FD. From Rogoff (1990) and Mani and Mukand (2007) we borrow the visibility concept in a two-politician model.

[4]This chapter is coauthored with Ruixin Wang.

Public goods are 'invisible' if they do not provide a good indication of politicians' capability, either because they are difficult to observe or because they are determined by factors beyond the government's control. Mani and Mukand (2007) showed that a government tends to spend more on visible projects than invisible projects, since voters infer its capability from visible projects. This is referred to as the visibility effect. We extend their model by introducing yardstick competition between two politicians. Using a one-shot game, we show that yardstick competition motivates the politicians to start a dress-up contest. To win more support in an election, they allocate more resources (public expenditure and effort) to the more visible goods, since these goods demonstrate their capability and provide a better image, given a binding fiscal budget constraint. In this sense, the yardstick competition turns into a competition for a better image, and FD can intensify this dress-up contest. Our model is related to the tax competition model (see for example Janeba and Peters (1999), Cai and Treisman (2005), and Zissimos and Wooders (2008)), but our conclusion is rather different. In the tax-competition literature, the mobility of capital motivates governments to promote public services. Using a similar framework, we show that the mobility of information may not always be positive, because it can distort the structure of public expenditure and lead to a welfare loss.

Second, we provide strong empirical evidence for public-expenditure distortion in visible and invisible goods, and we also find that this distortion caused by FD can result in a social welfare loss. To the best of our knowledge, although the visibility effect has been theoretically established, no research has empirically verified this effect, possibly because of the difficulty of finding good empirical proxies. In this paper, we investigate the FD effect on the regional poverty rate, an important aspect of social welfare. We propose to use cash assistance to the poor as a proxy for the more visible project, and vendor payments as a proxy for the less visible project. Using U.S. state level data from 1992 to 2008, we find that FD causes a public expenditure flow from the more visible to the less visible project. This result provides evidence for the visibility effect, and also confirms our theoretical findings that FD can cause dress-up contests between local governments. To further investigate the role of yardstick competition, we propose two yardstick competition measures based on the comparability of jurisdictions and the competitiveness of local governments. We estimate a difference-in-difference model and find that a stronger yardstick competition leads to a stronger FD effect on the structure of public expenditure. This is another evidence of dress-up contests. To capture how this distortion of public expenditure affects poverty, we use a functional coefficient approach, and we estimate a pooled panel and a panel with a fixed effect. This approach allows us to capture the possible nonlinear interaction between the cash-vendor-payment (CV) ratio, welfare expenditure, and poverty. We find that the distortion of public expenditure, measured by the CV ratio, can greatly weaken the effect of welfare expenditure on poverty reduction, and this influence appears to be nonlinear. Considering the possible endogeneity of welfare expenditure, we propose using public expenditure on health and hospitals as an instrumental variable of welfare expenditure. Our analysis shows that this instrument is valid theoretically and statistically. We thus empirically verify our theoretical findings, and provide empirical evidence for a dark side to FD.

The remainder of the paper is organized as follows. In the next section we formally model the causes and effects of dress-up contests in the presence of FD. Section 5.3 provides empirical evidence for dress-up contests, and Sections 5.4 and 5.5 analyse the FD effect on social welfare. Section 5.6 summarises and concludes.

5.2. The basic model

The basic model aims to illustrate how yardstick competition (dress-up contests), which is introduced by FD, can affect politicians' resource allocation to two types of public goods, more visible and less visible goods. In practice, yardstick competition can arise in two cases. First, two local politicians from the different (neighbouring) cities/towns are ambitious for one position in the higher level government. Second, each local incumbent politician competes with rivals to win his local election. If the rivals are cheap talkers, then the competition pressure comes from the neighbouring jurisdictions, and it is the performance of neighbouring politicians that help voters evaluate the capability of the local incumbent. In both cases voters compare these two incumbents from the neighbouring jurisdictions. The two cases are equivalent in modeling, and thus we only discuss the first case for simplicity. Since voters can only infer politicians' capability from public services, politicians aim to establish a better image to win votes. However, an overemphasis on image building can cause an efficiency loss in welfare expenditure, and further hurt social welfare. In this section, we first derive the equilibrium of a one-shot game, and then analyse the comparative statics, i.e. the impact of FD on this equilibrium.

5.2.1. Politicians

We follow Cai and Treisman (2005) and assume that politicians are partially self-interested, caring both voters' welfare and private interest. To model such partially self-interested politicians, we first discuss an ideal social planner whose aim is to appropriately distribute public resources (expenditure on public goods) to maximize the utility of the "society". The utility function of the social planner is

$$\begin{aligned} U_P &= \sum_{j=1}^{J} \upsilon_j z_j - C_P\left(e_1, \ldots, e_J\right), \\ \text{s.t.} \quad I &= \sum_{j=1}^{J} e_j. \end{aligned} \tag{5.1}$$

In the utility function, z_j is the observed outcome of the public good j, e_j is the public expenditure on public good j with the budget constraint I, and $C_P(\cdot)$ is the social cost of all public expenditure with $C_P'(e) > 0$ and $C_P''(e) > 0$. The social planner cares all public goods with different weights υ_j. This is an ideal case. In practice, however, a politician cares not only the social utility (benevolent), but also his own utility (self-interested), i.e. wining the election. Assume that a politician put weight γ on the social utility, and $(1-\gamma)$ on his own utility, and he maximizes his expected payoff function

$$\begin{aligned} \max_{e_{1i},\ldots,e_{Ji}} \mathrm{E}(U_i) &= \gamma U_{P,i} + (1-\gamma)\, R\eta_i - C_i\left(e_{1i}, \ldots, e_{Ji}\right) \qquad i \in \{A, B\}, \\ \text{s.t.} \quad I_i &= \sum_{j=1}^{J} e_{ji}, \end{aligned} \tag{5.2}$$

where R is the return from winning the election, with $R = 0$ indicating failure. η_i is the probability of winning the local election for the politician i. To win the future election, each politician needs to provides evidence of his capability (such as public services) at the cost C to convince voters. Note that C also depends on the public expenditure on J public services, but C is different from C_p, representing the extra cost for the politician to provide public goods besides social cost, e.g. management expenses, time, et al. We assume that the first- and second-order derivatives of the cost function satisfy $C(e)' > 0$ and $C''(e) > 0$. The budget constraint also applies.

5.2.2. Voters

We assume that there are two types of voters: well-informed voters (proportion k) and ill-informed voters (proportion $1-k$). Well-informed voters have their own ideology, i.e. their political persuasions and preference on governmental behavior (for example, more emphasis on defense or economic construction), and they make voting decisions based on the (inferred) capability of politicians and their ideology. If a politician's political persuasion is far away from a voter's ideology, then the voter is less likely to vote for this politician. Let s be the measure of a voter's ideology which uniquely identifies every voter. We assume that s is uniformly distributed between $[0, 1]$. Then the choice of voter s depends on

$$\pi_{s,i} = \frac{\Phi_i}{D_{s,i}} \qquad i \in \{A, B\}, \tag{5.3}$$

where Φ_i is the inferred capability of politician i, $D_{s,i}$ is the difference between the voter's ideology and politician i's political persuasions. More particularly, voter s chooses to support politician A if $\pi_{s,A} > \pi_{s,B}$ and vice versa. If we assume, without loss of generality, that politician A is the left wing, and B is the right, then $D_{s,A} = s$, and $D_{s,B} = 1 - s$. Given the inferred capabilities of the two politicians (which we shall discuss in details in the next subsection), we can compute the position of the marginal voter $\hat{s}$ that is indifference to politician A and B, that is $\pi_{s,\,A} = \pi_{s,\,B}$. This leads to an indifference marginal voter

$$\hat{s} = \frac{\Phi_A}{\Phi_A + \Phi_B} \tag{5.4}$$

This threshold value $\hat{s}$ also determines the share of well-informed voters supporting A and B. A simple calculation shows that well-informed voters with $s < \hat{s}$ will support A, while those with $s > \hat{s}$ will support B, i.e.

$$S_A = \hat{s} = \frac{\Phi_A}{\Phi_A + \Phi_B} \qquad S_B = 1 - \hat{s} = \frac{\Phi_B}{\Phi_A + \Phi_B}. \tag{5.5}$$

The ill-informed voters do not have ideology, and they make decisions randomly. We assume that politicians A and B equally share the votes of ill-informed voters. Then the probability of politician i to win election can be written as

$$\eta_i = kS_i + \frac{(1-k)}{2} \qquad i \in \{A, B\}, \tag{5.6}$$

and we always have $\eta_A + \eta_B = 1$.

5.2.3. Assessing politicians' capability

To model the dress-up contest, we consider two types of public goods: more visible goods a and less visible goods b. According to Mani and Mukand (2007), public goods are less visible if it is hard to assess governmental competence based on their observed outcome. Politicians need to allocate their limited resources to these two types of goods, and voters can then infer their capability. Following Mani and Mukand (2007), we assume that the production function of each good is

$$z_{j,i} = \tau_i + e_{j,i} + \epsilon_{j,i} \qquad j \in \{a, b\},\ i \in \{A, B\}, \tag{5.7}$$

where $z_{j,i}$ is the observed outcome of the public good j provided by politician i, τ_i is politician i's capability, $e_{j,i}$ is politician i's expenditure or effort on good j, and $\epsilon_{j,i} \sim N\left(0,\, \sigma^2_{j,i}\right)$ captures the exogenous stochastic factors. Public good a being more visible than b implies that there is more noise in the outcome of b than in that of a, i.e. $\sigma^2_{a,\,i} < \sigma^2_{b,\,i}$. Mani and Mukand (2007) provided two reasons for visibility differences. First, the outcome of some goods is intrinsically harder to directly observe or measure (e.g. those with short-term results are typically more visible than those that are long term). Second, some public goods are more 'complex' in the sense that their outcome is affected by a variety of factors other than governmental competence. For example, the quantity and quality of education is not determined only by governmental input, but also by teachers, parents, and peers. For simplicity and without loss of generality, politicians are assumed to have the same values of τ and ϵ_j.

Voters can observe the outcome of the public good z as well as the expenditure e. The politician's capability τ is unobserved, but voters have common knowledge of its prior distribution, $\tau_i \sim N\left(\overline{\tau},\, \sigma^2_\tau\right)$ for $i \in \{A, B\}$. Voters (with rational expectations) can use the observed outcome $z_i := \{z_{a,i}, z_{b,i}\}$ and the public expenditure $\mathbf{e}^*_\mathbf{i} := \{e^*_{a,i}, e^*_{b,i}\}$ to update their priors of the politicians' capability, i.e. from $\overline{\tau}$ to $(z_{j,i} - e^*_{j,i})$ with associated variance $\sigma^2_{j,i}$. According to Mani and Mukand (2007), the mean posterior assessment of the politician's capability can be obtained via

$$\Phi_i = \mathrm{E}\left(\tau_i \mid \mathbf{z}_i,\, \mathbf{e}^*_i\right) = \left[\frac{h_\tau \overline{\tau} + h_a\left(z_{a,i} - e^*_{a,i}\right) + h_b\left(z_{b,i} - e^*_{b,i}\right)}{h_\tau + h_a + h_b}\right],$$

where $h_\tau = 1/\sigma^2_\tau$ and $h_j = 1/\sigma^2_j$ $(j = a, b)$ are the precision of the prior and two realizations, respectively.

5.2.4. Equilibrium

Given the preference of voters, politician choose their strategies on $e_{a,\,i}$ and $e_{b,\,i}$ in a one-shot game simultaneously. We first look at the strategy of politician A. The optimization problem (5.2) gives the first order condition

$$\gamma \upsilon - \gamma C'_{P,\,A}\left(e_{a,\,A}\right) + (1-\gamma)\cdot Rk\cdot\left(\frac{h_a}{h_\tau + h_a + h_b}\right)\cdot\frac{\Phi_A}{\left(\Phi_A+\Phi_B\right)^2} - C'_A\left(e_{a,\,A}\right) - \lambda = 0$$

where λ is a Lagrangian multiplier. Since we have assumed the budget constraint is binding, λ must not be equal to zero, and the optimal expenditure $e^{\star}_{a,\,A} = \arg\max\left\{E\left(U_A\right)\right\}$. The case for politician B is symmetric.

5.2.5. Comparative statics

Based on the analysis above, we can examine how FD affects the politicians' behaviour, i.e. their public expenditure on the two types of goods, $e_{a,i}$ and $e_{b,i}$. FD can be regarded as a trigger of yardstick competition, which strengthens the information externality, and gives the voters more knowledge of the politicians' capabilities. This thus increases the proportion of well-informed voters, i.e. k, because information externality facilitates voters to detect and compare the public services, and further increases the comparability between the politicians in the neighboring jurisdictions. Therefore, we analyse the effect of FD by investigating how an increase in k affects the equilibrium.

We study the behavior of politician A. Define $F_A := \partial\widehat{\mathrm{E}}(U_A)^*/\partial e_{a,A}$, and we have

$$F_A = \gamma \upsilon - \gamma C'_{P,\,A}\left(e_{a,\,A}\right) + (1-\gamma)\,Rk\cdot\left(\frac{h_a}{h_\tau + h_a + h_b}\right)\cdot\frac{\Phi_A}{\left(\Phi_A+\Phi_B\right)^2} - C'_A\left(e_{a,\,A}\right) - \lambda.$$

Note that we always have

$$\frac{\partial F_A}{\partial k} = (1-\gamma)\,R\cdot\left(\frac{h_a}{h_\tau + h_a + h_b}\right)\cdot\frac{\Phi_A}{\left(\Phi_A+\Phi_B\right)^2} > 0,$$

and

$$\frac{\partial F_A}{\partial e_{a,\,A}} = -\gamma C''_{P,\,A}\left(e_{a,\,A}\right) - 2\,(1-\gamma)\cdot Rk\cdot\left(\frac{h_a}{h_\tau + h_a + h_b}\right)^2\cdot\frac{\Phi_A}{\left(\Phi_A+\Phi_B\right)^3} - C''_A\left(e_{a,\,A}\right) < 0.$$

Therefore, using the implicit function theorem, we obtain the following inequality at equilibrium

$$\frac{\partial e_{a,A}}{\partial k} = -\left(\frac{\partial F_A}{\partial k}\right)\Big/\left(\frac{\partial F_A}{\partial e_{a,A}}\right) > 0. \tag{5.8}$$

This shows that as k increases, politician A spends more on more visible public goods. Given the binding budget constraint, the expenditure on less visible goods thus shrinks. The analysis of politician B is similar, and we also have $\partial e_{a,B}/\partial k > 0$.

To summarise, our model shows that when k increases (more well-informed voters), politicians tend to focus more on establishing a good image. Given a binding fiscal budget constraint, more visible goods are more efficient at demonstrating capability and establishing a good image. This explains why expenditure on more visible goods increases under FD. However, an overemphasis on visible goods can lead to a structural bias in the public expenditure, and thus hurt social welfare. This implies that politicians' focus on their image may have a negative effect on social welfare. We investigate these theoretical findings empirically in the following sections.

5.3. Evidence for a dress-up contest

Our empirical analysis has two goals. The first is to provide evidence for the association between FD and dress-up contests. Second, we ask how dress-up contests affect poverty, an important aspect of social welfare. We address the first issue in this section, and the second in the following two sections. We use U.S. state level data, and our sample covers 48 states excluding Alaska and Hawaii with the time span from 1992 to 2008.

A key issue is how to determine the more visible and less visible public goods. It is difficult to find a strictly visible public good in the real world because the outcome of most such goods is determined by a number of factors beyond the government's control and is difficult to observe or measure. We focus on poverty, and we consider cash assistance to be relatively visible and vendor payments to be less visible. Cash assistance directly increases citizens' disposable income and reduces poverty. Hence, its outcome, i.e. poverty reduction, can be observed in the short term, and it primarily depends on the government's expenditure on this service. In contrast, vendor payments (welfare expenditure excluding cash assistance) are made to private purveyors for medical care, burials, and other commodities. The outcome of these payments depends on a large number of factors beyond the government's control, such as the performance of other institutes, and it may not be obvious in the short term. To appreciate these two measures, we need to distinguish between two concepts: visibility and visuality. Public services are visible if their outcomes are affected by less noisy factors, while they are visual if their outcomes are easily observed by voters. Some of the items of vendor payment can be visual, such as activities of soup kitchen. However, they are still less visible than cash because the outcome of these payment depends on the performance of intermediate institute, i.e. soup kitchen. Therefore, it is reasonable to regard cash assistance as more visible and vendor payments as less visible.

We provide evidence to show the existence of a dress-up contest. Since it is difficult to exactly identify all the transmission channels, we use evidence from different sources to rule out possible alternative explanations.

5.3.1. FD effect on public-expenditure structure

We first consider a direct test for the causal effect of FD on dress-up contests. To outline our empirical strategy, we introduce some preliminary notation. Assume that state-level politicians spend $1/v_S$ of the state expenditure on visible projects, while local-level politicians spend $1/v_L$ of the local expenditure on such projects. Since yardstick competition is more fierce in local elections than in state elections, we have $v_S > v_L \geq 1$. If we let Γ be the total (state + local) public expenditure, and let D be the degree of FD, then the total expenditure on the more visible project (cash assistance) and that on the less visible project (vendor payments) are given by

$$\text{Cash} = \frac{\Gamma}{v_S} + \Gamma \mathrm{D}\left(\frac{1}{v_L} - \frac{1}{v_S}\right) \quad \text{and} \quad \text{Vendor} = \Gamma - \text{Cash} = \Gamma\left(1 - \frac{\mathrm{D}}{v_L} + \frac{\mathrm{D}-1}{v_S}\right).$$

The ratio of Cash to Vendor (hereafter the CV ratio) is

$$\text{RCV} = \frac{v_L + \mathrm{D}(v_S - v_L)}{v_L(v_S - 1) - \mathrm{D}(v_S - v_L)}.$$

Note that RCV is a monotonically increasing function of the degree of FD, i.e.

$$\frac{\partial \text{RCV}}{\partial \mathrm{D}} = \frac{v_S v_L (v_S - v_L)}{[(v_S - 1)v_L - \mathrm{D}(v_S - v_L)]^2} > 0. \tag{5.9}$$

Therefore, as the degree of FD increases, the total expenditure on cash and the CV ratio both increase. Inequality (5.9) thus allows us to empirically test the direct association between FD and dress-up contests.

To test this association, we consider the reduced-form model

$$\text{RCV}_{it} = \alpha_i + \kappa_0 + \kappa_1 \mathrm{D}_{it} + \kappa_2 \text{TWE}_{it} + \varepsilon_{it}, \tag{5.10}$$

where the subscript it denotes observation of the ith state ($i = 1, \ldots, N$) at year t ($t = 1, \ldots, T$), and α_i is the individual-specific effect. D represents the degree of FD, and we measure it by

$$D := \frac{\text{Local public expenditure}}{\text{Total public expenditure}},$$

where the local expenditure includes the expenditure of the county, city, and town governments, and the total expenditure is the expenditure of the state and local governments. TWE is the total (state + local) welfare expenditure. The fixed-effect estimation results[5] are given in column (1) of Table 5.1. It shows that a larger degree of FD is associated with a larger CV ratio, and the correlation is strong and robust. In columns (2) and (3), we replace the contemporary FD D by its first- and second-order lagged values D_{L1} and D_{L2}, respectively, to capture the causal effect, since an FD policy may take effect after a period of time. We see that using lagged values gives a more positive and more significant estimate, confirming the causal relationship between FD and the CV ratio.

Table 5.1: FD effect on CV ratio

	(1)	(2)	(3)	(4)
D	0.4849			
	(2.43)			
D_{L1}		0.5018		
		(2.67)		
D_{L2}			0.4888	
			(3.06)	
D_{CT}				2.3915
				(2.77)
TWE	−0.2464	−0.2539	−0.2270	−0.2760
	(−9.78)	(−10.26)	(−10.06)	(−5.68)
CONST	0.2430	0.2484	0.2233	0.2532
	(7.83)	(8.40)	(8.58)	(6.86)

As a robustness check, we recompute the FD ratio using the expenditure from only the city and town governments (excluding the county-level governments), and we denote this ratio D_{CT}. Since yardstick competition is supposed to be more intense at the city and town level, we expect to observe a stronger association between dress-up contests and FD, i.e. a more significant and positive estimated coefficient κ_1. The results in column (4) indeed indicate a more significant effect, showing the robustness of this finding.

[5] A preliminary analysis suggests the fixed-effect model is more appropriate than the random-effect model.

5.3.2. The role of yardstick competition

Yardstick competition plays a crucial role in the theoretical model, so we introduce it into our empirical analysis. Since FD distorts the structure of public expenditure through the channel of yardstick competition, we expect that the stronger the yardstick competition, the stronger the distortion. More formally, if the local-level yardstick competition is intensified, i.e. v_L is smaller, then the CV ratio increases, because

$$\frac{\partial \mathrm{RCV}}{\partial v_L} = -\frac{\mathrm{D}v_S^2}{\left[(v_S-1)v_L - \mathrm{D}(v_S - v_L)\right]^2} < 0.$$

This implies that given the same degree of FD, if the local-level yardstick competition is stronger in a particular state, then the politicians in that state have more incentive to invest in visible projects. In other words, the degree of yardstick competition can affect the impact of FD on the structure of public expenditure. This mechanism can be empirically captured by an interaction term between yardstick competition and FD. Thus, we consider the model

$$\mathrm{RCV}_{it} = \alpha_i + \kappa_0 + \kappa_1 \mathrm{D}_{it} + \kappa_2 \mathrm{COMP}_{it} + \kappa_3 \mathrm{D}_{it} \times \mathrm{COMP}_{it} + \kappa_4 \mathrm{TWE}_{it} + \varepsilon_{it}, \quad (5.11)$$

where COMP is a measure of the yardstick competition. Estimating (5.11) allows us to identify the mechanism described in Section 5.2, at least to some extent.

Yardstick competition is a difficult concept to measure, and to the best of our knowledge there is no satisfactory measure in the literature. We propose two measures based on the comparability of jurisdictions and the competitiveness of local governments. First, we consider the comparability of jurisdictions. This is motivated by the argument of Bodenstein and Ursprung (2005) that yardstick competition 'emerges when the performance of governments in various jurisdictions becomes sufficiently comparable so that the voters can alleviate the agency problem by making meaningful comparisons between jurisdictions'; see also Besley and Case (1995). In the U.S., most congressional districts consist of several local governments that have similar political and economic situations, such as similar political interests and voters' preferences. Hence, we expect that the yardstick competition between local governments *within* a congressional district is stronger than that *outside* the district. This implies that the congressional district demarcates the political boundaries of the yardstick competition. If a given district contains more local governments, then the yardstick competition in this district is more intense because each local government has more comparable rivals. Given this motivation, we propose to measure the yardstick competition by

$$\mathrm{COMP}_r := \frac{\text{Number of local governments}}{\text{Number of congressional districts}}.$$

This ratio is unaffected if we control for a state's land size or population since we divide both the numerator and denominator by the land size or population.

Next, we consider measuring the yardstick competition by the competitiveness of the local elections, which is computed based on the percentage of votes won by the leading party. We denote this measure as COMP_c. The average level of competitiveness is a reasonable measure of the yardstick competition within the state. The competitiveness is higher if the leading party wins a smaller share of the votes, suggesting that the competing parties are well matched or none of the candidates has strong support. In both cases, the yardstick competition can be intense. Due to the lack of county-level data, we use congressional-district data. In the two-party system of the U.S., congressional elections are expected to be highly correlated with local (county, city, or town) elections, and thus the average competitiveness of these elections can be a proxy for the yardstick competition at the local level.

To see how the FD effect varies at different levels of competitiveness, we first rank all the states according to their average competitiveness (averaged over time). Then, we estimate the FD effect using two samples, made up of the most competitive and the least competitive states. Columns (1)–(4) of Table 5.2 present the results. It is clear that the FD effect on the CV ratio is much stronger and more significant in the more competitive states. Next, we examine the interaction effect of competitiveness more formally by estimating the panel data model (5.11). The results are given in columns (5)–(8) of Table 5.2. We see that the interaction terms are strongly positive when using COMP_r and strongly negative when using COMP_c in the models with contemporary and lagged FD. This again confirms that a stronger yardstick competition leads to a stronger FD effect on the CV ratio. The significance of the level terms D and COMP depends on the measurement of the yardstick competition. COMP is significant but D is not when we measure the competition by COMP_r; and D is significant but COMP is not when we measure the competition by COMP_c. In the difference-in-difference model, the coefficients of the level terms capture only an 'initial' effect. The different significance levels suggest that COMP_r and COMP_c measure the yardstick competition from different perspectives. Since the size of the interaction term in columns (5) and (6) is much larger than that in columns (7) and (8), and is also larger than the size of its level terms, we find that the results from the two measures are generally consistent: a larger degree of FD and more intense yardstick competition are associated with a higher CV ratio.

Table 5.2: Interaction between FD, yardstick competition, and CV ratio

	15 most competitive		15 least competitive		Entire sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
D	0.7432	0.5789	−0.0900	0.4929	0.0599		0.8393	
	(4.91)	(2.52)	(−0.40)	(1.55)	(0.39)		(2.92)	
D_{L1}						0.1178		0.9182
						(0.79)		(3.60)
TWE	−0.2923	−0.2769	−0.2804	−0.2195	−0.2636	−0.2668	−0.2457	−0.2536
	(−8.00)	(−4.94)	(−8.76)	(−6.50)	(−11.90)	(−12.39)	(−10.24)	(−10.64)
$COMP_r$	√		√		−2.4494	−2.7143		
					(−3.56)	(−4.01)		
$COMP_c$		√		√			−0.0004	−0.0004
							(−0.66)	(−0.56)
$D\times COMP_r$					6.9800			
					(3.91)			
$D\times COMP_c$							−0.0067	
							(−2.24)	
$D_{L1} \times COMP_r$						6.3943		
						(4.05)		
$D_{L1} \times COMP_c$								−0.0081
								(−3.05)
CONST	0.2590	0.2591	0.3086	0.2328	0.3266	0.3335	0.2661	0.2722
	(8.07)	(3.68)	(8.62)	(8.74)	(11.50)	(12.54)	(5.93)	(5.90)

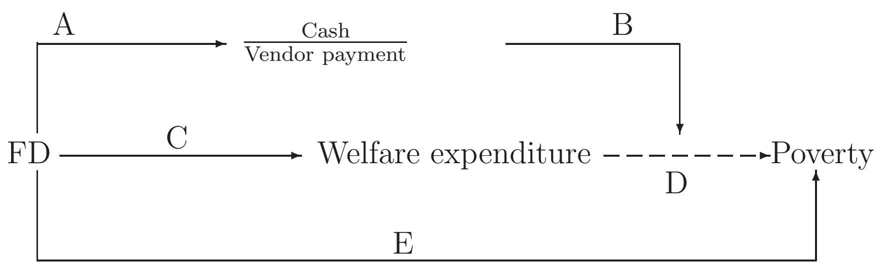
Note: Columns (1) and (2) use the 15 most competitive states, based on $COMP_r$ and $COMP_c$, respectively; columns (3) and (4) use the 15 least competitive states, based on $COMP_r$ and $COMP_c$, respectively; columns (5) and (6) use the entire sample of 48 states.

To summarise, the above analysis shows that a high degree of FD is associated with an expenditure flow from the more visible product (cash assistance) to the less visible product (vendor payments), and the association is even stronger in regions with more intense yardstick competition. This is because to achieve a better image and win more votes, politicians tend to allocate more resources to the more visible project. This dress-up contest is intensified by FD through the channel of yardstick competition. These empirical results thus provide support for our theoretical findings.

5.4. FD effect on poverty

We have seen, from both theoretical and empirical perspectives, that fiscal FD can cause a dress-up contest which forces governments to allocate more expenditure to the more visible public goods. In the following two sections, we investigate how this distortion of public expenditure influences social welfare. We focus on the effect of FD on the poverty rate, an important aspect of social welfare, and empirically identify the transmission mechanisms. For this purpose, we introduce three additional variables: poverty (p), unemployment rate (UNEM), and Gini index (GINI). Poverty is defined by the share of people with an income lower than the standard income, and this standard differs across states. A more detailed description of the variables and their sources is given in the Appendix.

Figure 5.1: Transmission channels from FD to poverty



We focus on three channels from FD to poverty, which are summarised in Figure 5.1. First, according to the two-politician model in Section 5.2, FD can affect poverty through the dress-up contest, i.e. an expenditure flow from less visible goods to more visible goods (effects A and B). Second, FD can indirectly affect poverty by affecting the welfare expenditure (effects C and D). On the one hand, FD may increase the welfare expenditure due to higher administrative costs; on the other hand, it is likely that welfare expenditure shrinks after FD because the mobility of the poor motivates governments to spend less on welfare to reduce the fiscal burden. It is not clear which effect dominates, and we investigate this in our empirical study. Finally, in addition to the indirect effects, FD can have an impact on poverty through channels other than welfare expenditure and dress-up contests. We consider other connections between FD and poverty as effect E.

We observe that the CV ratio influences poverty not directly but indirectly, by changing the structure of welfare expenditure. Hence, the arrow line of effect B does not point at poverty but at effect D. We use a dashed line for channel D since there is potential reverse causality between welfare expenditure and poverty, which we will investigate using instrumental variables.

5.4.1. Standard panel data

To provide empirical evidence for the transmission channels described in Figure 5.1, we first identify each effect A–E separately. First, we examine the transmission channel from FD to welfare expenditure, and then to poverty, namely effects C and D. To show the mediation of the welfare expenditure, we estimate the following models:

$$\mathrm{TWE}_{it} = \alpha_i + \theta_0 + \theta_1 \mathrm{D}_{it} + e_{it}, \tag{5.12}$$

$$p_{it} = \alpha_i + \beta_0 + \beta_1 \mathrm{D}_{it} + \beta_2 \mathrm{TWE}_{it} + \beta_3 \mathrm{UNEM}_{it} + \beta_4 \mathrm{GINI}_{it} + \epsilon_{it}. \tag{5.13}$$

Model (5.12) captures the transmission effect C, while (5.13) captures the direct effect of FD on poverty (effect E) and the indirect effect through welfare expenditure (effect D).

Table 5.3: Results for separate transmission channels

	(1) TWE	(2) TWE	(3) p	(4) p	(5) p	(6) p	(7) p
D	−1.1031		7.4507	5.4731	4.7246	3.3062	4.6255
	(−3.58)		(4.54)	(3.49)	(2.29)	(1.54)	(1.82)
D_{L1}		−1.0418					
		(−3.56)					
TWE				−1.7927	−2.2731	−0.4419	−2.3902
				(−3.38)	(−3.57)	(−0.72)	(−3.48)
GINI					−0.1913		0.6650
					(−0.08)		(0.29)
UNEM					0.5519		0.5784
					(6.07)		(5.78)
RCV						4.5906	0.1824
						(3.19)	(0.13)
CONST	0.7922	0.7796	12.037	13.457	10.960	12.142	10.518
	(28.15)	(27.82)	(80.16)	(30.95)	(8.46)	(22.83)	(8.00)

Columns (1)–(5) of Table 5.3 present the standard fixed-effect estimation results based on Equations (5.12) and (5.13). Column (1) shows that FD has a strongly negative effect on welfare expenditure. Column (2) replaces the contemporary value of FD by its first-order lagged value D_{L1}, and shows a similar result, confirming that a high degree of FD leads to less welfare expenditure. This suggests that the negative effect of FD on welfare expenditure dominates in our case. In particular, since the poor are mobile, an increase of welfare expenditure in one jurisdiction attracts the poor to this region, which adds to the burden of this jurisdiction but reduces the burden of others. Therefore, if most jurisdictions are free riders, then FD leads to a coordination failure and the inefficient provision of public goods. Column (3) shows a significant and positive overall effect of FD on poverty, challenging the conventional viewpoint that it has a positive impact on social welfare. This effect is largely reduced (in size and significance) when we include welfare expenditure (column (4)), but remains strong, and the coefficient of welfare expenditure is significantly negative. This suggests that part of the FD effect on poverty is explained by the intermediate transmission through welfare expenditure, and it provides evidence for strong effects C and D. These effects are robust when we include the Gini coefficient and unemployment (column (5)).

To examine effect B, we first add RCV as an explanatory variable in the poverty regression. Columns (6) and (7) show that the FD effect remains strong and positive after we control for welfare expenditure and the CV ratio, and this suggests the existence of effect E. The strongly positive and robust effect of FD again confirms the negative effect of FD on poverty reduction. The CV ratio is positively related to poverty, but this effect becomes insignificant when we control for unemployment and the Gini index. It shows that the CV ratio can be positively related to poverty, but the delicate coefficient suggests that the standard panel data model may not fully capture the effect of the CV ratio on poverty. Also, we see that including RCV can affect the estimated coefficient of WE, which suggests possible interactions between RCV and WE. In fact, the CV ratio influences poverty by interacting with the effect of welfare expenditure. An excessively large (or small) CV ratio reduces the effect of welfare expenditure on poverty reduction, while an appropriate value of the ratio can maximise the effect of welfare expenditure. Therefore, effect B cannot be fully captured by the standard fixed-effect model with RCV as a control variable, and more appropriate methods are required.

5.4.2. Endogeneity of welfare expenditure

A potential issue is the endogeneity of welfare expenditure. The endogeneity is due to possible reverse causality between welfare expenditure and poverty; in particular, welfare expenditure can reduce poverty, while regions with a higher poverty rate are likely to have more welfare expenditure. To reduce the potential bias caused by reverse causality, we consider instrumental-variable estimation. We propose to use public expenditure on health and/or hospitals as the instrumental variable of welfare expenditure. Expenditure on health and hospitals is highly correlated with welfare expenditure because factors such as citizens' interest in government services, politicians' attention to citizens' wellbeing, and the power of the public-sector unions can influence the expenditure on welfare, health, and hospitals. Moreover, this instrument does not depend on poverty because government hardly increase or reduce health expenditure for poverty reason. Also, there is no clear transmission channels from public health expenditure on poverty other than welfare expenditure. This is because only a proportion of public expenditure on health and hospital may be distributed to *individual* health care, and only this part of expenditure is possibly related with poverty. Even if part of public expenditure on health and hospital is related with poverty, it is still unclear how much assistance to individual health expenditure can alleviate poverty.Therefore, health and hospital expenditure satisfies the requirements of relevance and exogeneity, so it is an appropriate instrumental variable. This instrument is in the similar spirit of Levitt (2002) who used expenditure on fire fighting as an instrument of expenditure on police when investigating the determinants of crime.

Table 5.4: Results of poverty regression: IV estimation

	(1)	(2)	(3)	(4)	(5)	(6)
Instrument	HE	HO	HH	HH	HEO	HEO
D	3.5819	5.1961	4.4651	4.6046	4.0719	4.5837
	(1.90)	(2.56)	(2.44)	(2.44)	(2.50)	(2.54)
TWE	−3.5619	−1.7413	−2.5657	−3.2093	−2.9985	−3.5670
	(−2.87)	(−1.14)	(−2.25)	(−1.73)	(−3.05)	(−2.06)
GINI	0.9078	−0.6449	0.0582	0.7794	0.4733	0.7590
	(0.35)	(−0.24)	(0.02)	(0.31)	(0.18)	(0.30)
UNEM	0.5360	0.5585	0.5483	0.5943	0.5424	0.6001
	(8.75)	(9.01)	(9.03)	(7.79)	(9.12)	(7.64)
RCV				−0.9291		−1.4686
				(−0.33)		(−0.55)
First-stage F-stat.	122.13	75.64	146.95	76.61	80.98	38.27
p-value of first-stage F-test	0.00	0.00	0.00	0.00	0.00	0.00
p-value of Hansen's J-test					0.24	0.24

Note: The dependent variable in all models is poverty. Columns (1)–(4) are 2SLS, and columns (5) and (6) are GMM. Using 2SLS to estimate columns (5) and (6) leads to consistent results.

We consider four variations of the instrument: expenditure on health (HE), expenditure on hospitals (HO), expenditure on health and hospitals (HH), and expenditure on health together with expenditure on hospitals (HEO). In the first three cases, the model is exactly identified, and we estimate it using two-stage least squares (2SLS). In the last case, we estimate the overidentified model using the generalised method of moments (GMM). The results are presented in Table 5.4. We see that using the instrumental variable does not change our results. In particular, the estimated coefficient of welfare expenditure using 2SLS/GMM remains significantly negative, and it is slightly larger (in absolute value) than the standard fixed-effect coefficient estimate, except in column (2). The estimates of the other covariates are generally unaffected when we use 2SLS. The first-stage F statistic and its p-value show that the instruments are in general highly correlated with the endogenous variable. However, the single-instrument HO is relatively weak compared with HE, and this explains the small absolute value of the welfare-expenditure coefficient in column (2). In columns (5) and (6), the rejection of Hansen's J test suggests that the overidentified instruments satisfy the orthogonal conditions, and thus are valid instruments.

To conclude, the results from a separate estimation of each channel show that effects A–E indeed exist. FD can have an impact on poverty by reducing welfare expenditure and, more interestingly, through the CV ratio. However, we also note that the interaction between poverty and the CV ratio (effect B) cannot be fully captured by the standard panel data model, and more thorough studies will be required.

5.5. Joint estimation using functional coefficient model

The above analysis specifies each channel separately, and shows that each effect is strong and significant. However, these channels may not be jointly strong and their relative importance is not yet clear. For example, it is possible that the transmission channel through the CV ratio (effects A and B) is individually significant, but plays a minor role when we control the channel through the welfare expenditure. Also, the standard fixed-effect model considered in the previous section cannot capture the interaction between the CV ratio, welfare expenditure, and poverty. A frequently used method to capture the interaction effect is difference-in-difference estimation:

$$p_{it} = \alpha_i + \beta_0 + \beta_1 \mathrm{D}_{it} + \beta_2 \mathrm{TWE}_{it} + \beta_4 \mathrm{RCV}_{it} + \beta_5 \mathrm{TWE}_{it} \times \mathrm{RCV}_{it} + \sum_{k=1}^{2} \gamma_k x_{it,k} + \epsilon_{it}, \quad (5.14)$$

where $x_{it} = (\mathrm{GINI}_{it}, \mathrm{UNEM}_{it})$. We argue that this approach does not work here, for two reasons. First, since RCV is influenced by D, the interaction term TWE $\times$ RCV can be highly correlated with the level terms even if all the variables are centred to remove multicollinearity, and therefore the estimated coefficient of the interaction term can be inefficient. Second, the interaction term provides only a positive or negative (linear) interaction effect, and this effect is the same for all CV ratio levels. However, it is possible that the welfare-expenditure effect on poverty depends *nonlinearly* on the CV ratio; in particular, both extremely large and small values of the CV ratio reflect the distortion of welfare expenditure, and this distortion can weaken its effect on poverty reduction. Therefore, the welfare-expenditure effect is expected to be a nonlinear function of the CV ratio (roughly U-shaped). This nonlinear relationship cannot be captured by Equation (5.14). Indeed, estimates of Equation (5.14) show that $\hat{\beta}_5$ is not significant.

5.5.1. Standard functional coefficient model

To investigate the relative importance of each channel and capture the possibly nonlinear relationship between the CV ratio and poverty, we consider the functional coefficient model in which the slope coefficients are allowed to vary over a common variable. We first consider a standard functional coefficient model,

$$p_{it} = \delta_0 + \delta_1 \mathrm{D}_{it} + \delta_2 \mathrm{TWE}_{it} + \delta_3 \mathrm{GINI}_{it} + \delta_4 \mathrm{UNEM}_{it} + \eta_{it}, \quad (5.15)$$

where the slope coefficient δ_k $(k = 0, 1, \ldots, 4)$ is a continuous function of the CV ratio. The variables D, TWE, GINI, and UNEM in Equation (5.15) are the same as in Equation (5.13), except that DINC is not included to avoid possible multicollinearity between TWE and DINC. Our robustness check suggests that including DINC does not change the shape of the curves, but just widens the confidence bands. One advantage of a functional coefficient model is that it allows regressors to be correlated with the smoothing variable RCV, and thus avoids the multicollinearity problem in (5.14). Moreover, it provides information on how the effect of welfare expenditure varies (possibly nonlinearly) for different values of the CV ratio. The model also allows us to rule out other possible transmission channels from the CV ratio to poverty, at least to some extent, if the other functional coefficients (δ_1, δ_3, and δ_4) do not vary over RCV or show no clear trends. For the moment, we consider a standard functional model without an individual-specific effect α_i (pool estimation), and the estimated coefficients are consistent if α_i is assumed to be uncorrelated with the regressors. In the next subsection we will allow correlation between α_i and the regressors and estimate a fixed-effect functional coefficient model.

The parameters in this model are estimated by local linear estimation (Fan and Gijbels (1996); see also Cai et al. (2000)). Thus we specify

$$\delta_k = \delta_{Ck} + \delta_{Sk}(\mathrm{RCV} - u_0) \quad (k = 0, 1, \ldots, 4) \tag{5.16}$$

where $\min(\mathrm{RCV}) \leq u_0 \leq \max(\mathrm{RCV})$. The parameters $(\delta_{Ck}, \delta_{Sk})$ are estimated by minimising the following objective function:

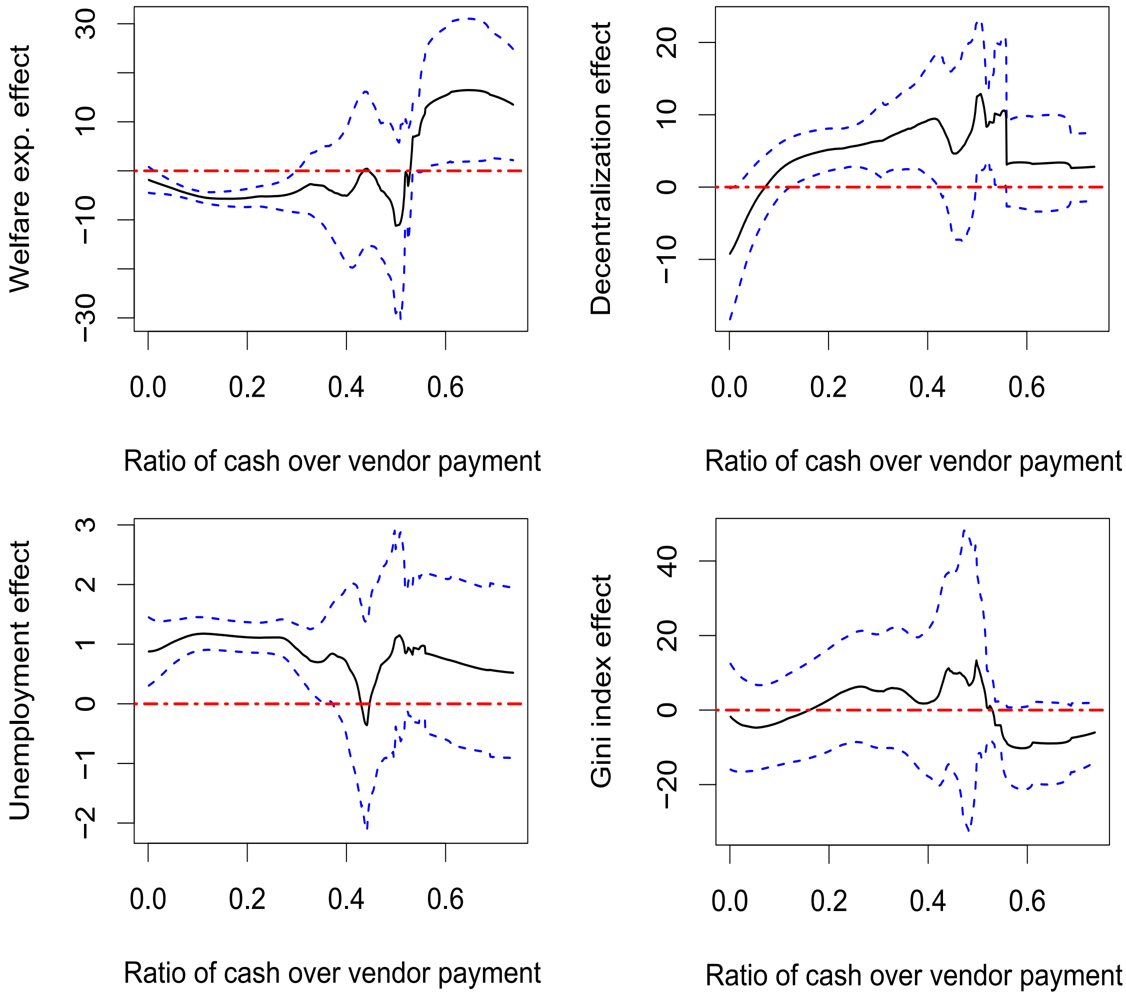
$$\min_{\delta_{Ck},\delta_{Sk}} \sum_i \sum_t \left(p_{it} - \sum_{k=0}^{4} \{\delta_{Ck} + \delta_{Sk}(\mathrm{RCV}_{it} - u_0)\} x_{itk} \right)^2 K_h(\mathrm{RCV}_{it} - u_0),$$

where x_{itk} is the kth regressor, and $K_h(\cdot) := h^{-1}K(\cdot/h)$ with bandwidth h and kernel function $K(\cdot)$. Various data-driven methods could be used to select the bandwidth, e.g. cross-validation (Fan and Gijbels, 1996). We choose the bandwidth by minimising the averaged mean square error, following Cai et al. (2000).

Figure 5.2 shows the slope parameters as a function of the CV ratio. The solid line plots the coefficient estimate, and the dashed lines are $\pm 2 \times$ the bootstrap standard errors (calculated over 200 replications). We see a rough U-shape of the welfare-expenditure effect on poverty (upper-left subfigure). The effect is significantly negative when the proporition of the cash assistance is relatively small, and it becomes stronger (more negative) as the ratio increases to around 0.2. However, when the ratio is more than 0.3, increasing the cash proportion weakens the welfare-expenditure effect on poverty reduction, with wide confidence bands. The effect even becomes weakly positive when the ratio is particularly high. The nonlinear behaviour shows that a deviation of the CV ratio from its optimal value, and in particular an increase in its value, can weaken the poverty-reduction effect of welfare expenditure. This provides evidence for the efficiency loss caused by an overemphasis on visible products.

The FD effect on poverty (upper-right subfigure) is significantly positive for values of the CV ratio from around 0.1 to 0.4, and less significant for larger values. The estimated

Figure 5.2: Marginal effect of control variables on poverty as function of CV ratio (standard functional coefficient model)



functional coefficients of the welfare expenditure and FD confirm the results from the standard fixed-effect model that the indirect channel (effects C and D) is strong, other channels also matter (effect E), but the evidence for the direct effects (A and B) is not as clear. We also see that the curves of FD, unemployment, and the Gini index have no particular shape, suggesting that the CV ratio does not influence poverty through these channels.

5.5.2. Fixed-effect functional coefficient model

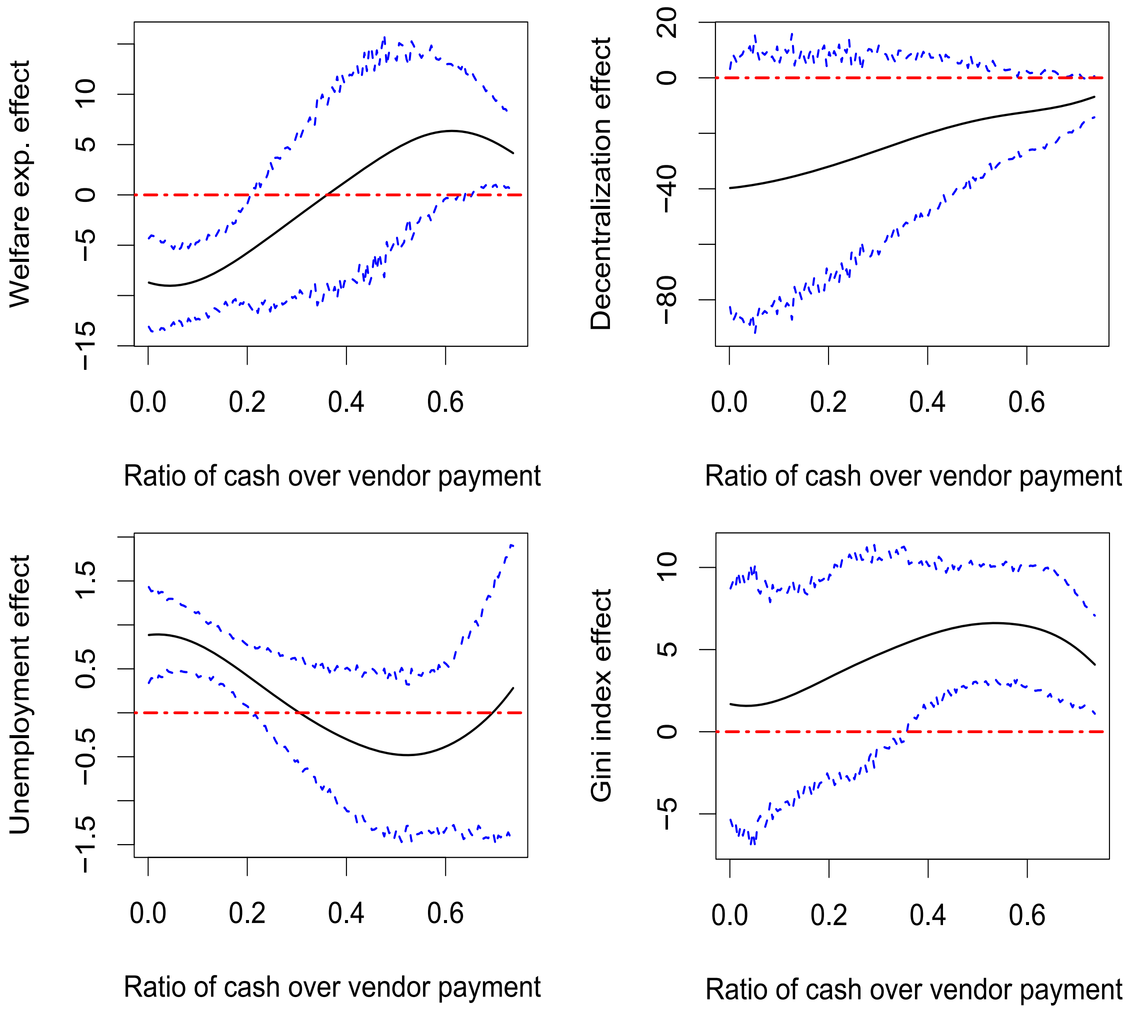
Standard functional coefficient estimation works if the individual-specific effect α_i is independent of the control variables. However, it is possible that an unobserved individual effect α_i is correlated with the control variables, for example, the historical and cultural differences between states (an unobserved individual effect) may affect the government behaviour, and thus impact the degree of FD. To allow for possible correlation between the individual-specific effect and the regressors, we estimate a fixed-effect functional coefficient model:

$$p_{it} = \alpha_i + \delta_0 + \delta_1 \mathrm{D}_{it} + \delta_2 \mathrm{TWE}_{it} + \delta_3 \mathrm{GINI}_{it} + \delta_4 \mathrm{UNEM}_{it} + \eta_{it}, \tag{5.17}$$

where α_i can be correlated with the regressors in any (unknown) pattern. In a functional coefficient model, the fixed effect cannot be removed by a preliminary step of first-difference or within-transformation of the dependent and independent variables, because the slope coefficients $\delta_k = \delta_k(\mathrm{RCV}_{it})$ are no longer constant for all the observations. The transformation based on equations also does not work, because it involves an additive function that impedes kernel-based estimation, and also because it produces an inconsistent estimated coefficient of the time-invariant term (see Sun et al. (2009) for the details). Therefore, we follow Sun et al. (2009) and remove the fixed effects by deducting a smoothed version of the cross-time average from each individual unit. This approach first analytically finds the fixed-effect vector via a weighted least square dummy variable model, and then estimates the functional parameters nonparametrically using a concentrate weighted least square method. To calculate the bootstrap standard error in the panel data model, we follow Kapetanios (2008) and construct bootstrap samples by resampling whole cross-sectional units with replacement (cross-sectional resampling).

Figure 5.3 presents the fixed-effect functional coefficient estimates for each control variable. In general the shape of the curves is similar to those in the standard func-

Figure 5.3: Marginal effect of control variables on poverty as a function of CV ratio: Fixed-effect estimation



tional coefficient model. In particular, the trends of the welfare-expenditure effect are consistent: welfare expenditure has a significantly negative effect on poverty when the CV ratio is low, but a weakly positive effect when the ratio is high (the U-shaped curve). Also, this effect becomes less significant as the ratio increases. The estimated coefficients of FD are below the zero line; they are much lower than those in the standard panel data model, even though we observe only the upper bound of the confidence interval. Thus, FD has little impact on the poverty rate if we control for the size (effects C and D) and the structure (effects A and B) of welfare expenditure. The results for unemployment and the Gini index show no particular trends.

In contrast to the standard estimation, the fixed-effect estimation results provide evidence for both the direct channels (effects A and B) and indirect channels (effects C and D), while the other channels (effect E) become relatively weak. In general, the harmful effect of FD can be observed in the functional coefficient analysis when we take the fixed effect into consideration.

5.5.3. Robustness check

We investigate the robustness of our results in various ways. First, we focus on the coefficient of TWE and consider different subsets of auxiliary variables {D, UNEM, GINI}. The results from both the standard and fixed-effect models show that including different auxiliary variables does not affect the curves of the welfare-expenditure effect.

Second, we consider using an alternative data set, namely the local governments' expenditure on cash and vendor payments. To ensure that the ratio is well-defined, we assign zero to those observations with no such assistance or payments.[6] We estimate the functional coefficient model using the local government expenditure. The left panel of Figure 5.4 shows that the welfare-expenditure effect on poverty is negative when RCV is small but weakly positive when RCV is large. This result is consistent with our previous findings. The larger confidence bands for small values of RCV are partly because we assign zero to those observations with no assistance or payments, which reduces the accuracy.

Finally, we consider the possible effect of lagged variables. This captures the causal effect, and using the lagged value can also reduce the endogeneity to some extent. We

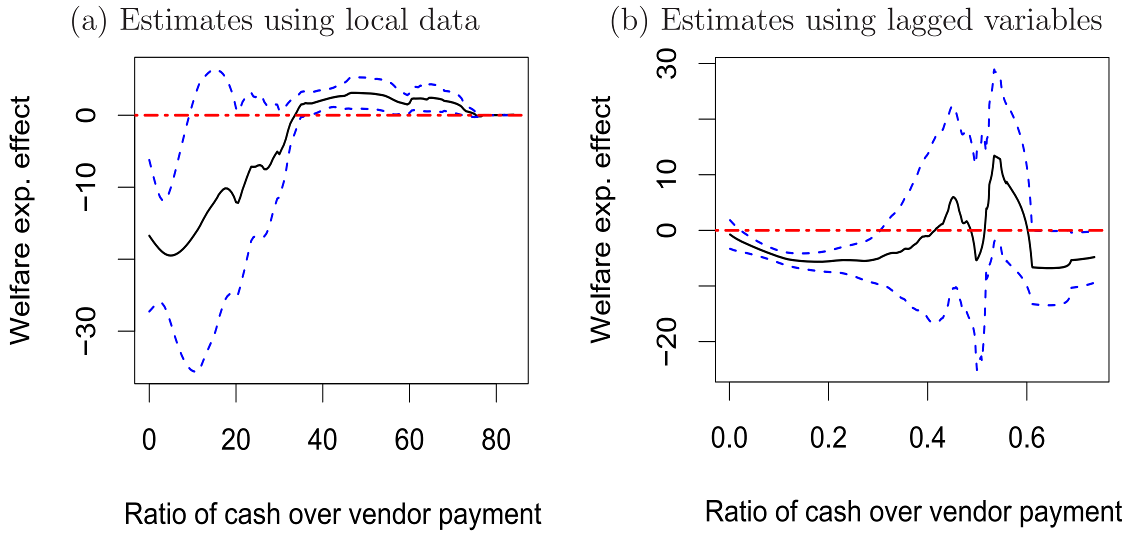
[6] Setting these observations to zero cannot distinguish the case with no cash and vendor payments from the case with vendor payments but no cash assistance.

consider the following model:

$$p_{it} = \alpha_i + \delta_0 + \delta_1 \mathrm{D}_{i,t-1} + \delta_2 \mathrm{TWE}_{i,t-1} + \delta_3 \mathrm{GINI}_{it} + \delta_4 \mathrm{UNEM}_{it} + \eta_{it}, \tag{5.18}$$

where δ_k is a function of $\mathrm{RCV}_{i,t-1}$. In this model, we take a first-order lag of the control variables D and TWE together with the smoothing covariate RCV, because they are related to the fiscal policies. We estimate (5.18) using both standard and fixed-effect models, and the right panel of Figure 5.4 shows that our main results are not affected.

Figure 5.4: Functional coefficient estimates of welfare-expenditure effect: Robustness check



5.6. Concluding remarks

This paper models and empirically identifies the dress-up contest introduced by FD and its harmful effect on social welfare. Because of asymmetric information, voters cannot observe politicians' capabilities, but they make assessments based on the outcome of public projects. Therefore politicians, under election pressures, are motivated to allocate more resources to more visible projects to improve their image. We show that the yardstick competition triggered by FD can turn into a competition for a better image, i.e. a dress-up contest, and this contest further causes a structural bias in public expenditure (more expenditure on visible projects) and reduces the efficiency of public expenditure.

Our empirical analysis first examined each transmission channel separately using the standard panel data model, and found that each effect is individually strong. On the one hand, FD significantly reduces the welfare expenditure, and thus further increases poverty. On the other hand, it encourages governments to spend more on visible projects, leading to a higher CV ratio in welfare expenditure. To capture the possible nonlinear interaction between CV ratio, welfare expenditure, and poverty, and also to examine the relative importance of each channel, we estimated the effects jointly using the functional coefficient panel data model. It showed that the transmission effects through welfare expenditure and the CV ratio are both nontrivial. An excessively large CV ratio weakens the poverty-reduction effect of welfare expenditure because of the efficiency loss. Separate estimation and joint estimation together provide supporting evidence for the positive effect of FD on poverty, and our results are robust to different model specifications. Therefore, our empirical analysis suggests that FD in general has a dark side that can lead to a higher level of poverty through a dress-up contest.

Our main results have important policy implications. Policymakers, who consider FD to be an efficient policy tool, should also be aware of its dark side. Two methods can help to avoid dress-up contests and their negative effects on social welfare in the course of FD. First, there should be a minimum level of public expenditure on less visible projects, so that the structure of public spending does not become too distorted. Second, an evaluation system could be introduced to increase the visibility of public projects, such as the CPA (comprehensive performance assessment) system used in the UK since 2002. Such an assessment system would allow voters to better evaluate politicians' capabilities.

Further research is needed in several areas. First, we plan to use an alternative measure of yardstick competition to provide further evidence for dress-up contests. Second, there are missing values in the current data set, and a better data set is thus required. Finally, we will consider functional coefficient estimation in the presence of instrumental variables.

Appendix

Detailed data description

p Regional poverty rate. The share of people with an income lower than the regional standard (varying over regions), from 1992 to 2008. *Source: Statistical Abstract.*

D Fiscal decentralisation. Defined as

$$\text{D} = \frac{\text{local expenditure}}{\text{local expenditure} + \text{state expenditure}},$$

where state expenditure refers to the expenditure by the state government, and local expenditure is the expenditure by all local governments. *Source: Statistical Abstract.*

UNEM Unemployment rate, from 1992 to 2008. *Source: Bureau of Economic Analysis.*

COMP$_r$ Yardstick competition based on the comparability of jurisdictions, the ratio of the number of local governments over the number of congressional districts.

COMP$_c$ Yardstick competition based on the competitiveness of local governments, the percentage of votes won by the leading party, from 1992 to 2008.

TWE Welfare expenditure. Total public welfare per thousand persons measured in thousands of dollars, from 1992 to 2008 excluding 2000, 2002, and 2006. *Source: Statistical Abstract.*

RCV The ratio of cash assistance to vendor payments. In this ratio, cash assistance is paid directly to needy persons under the categorical programs (Old Age Assistance, Temporary Assistance for Needy Families (TANF)) and any other welfare programs. Vendor payments are made directly to private purveyors for medical care, burials, and other commodities and services provided under welfare programs; and for the provision and operation by the government of welfare institutions. *Source: Statistical Abstract.*

GINI Gini index. *Source: Statistical Abstract.*

Chapter 6

Supplementary document to Chapter 2

6.1. Introduction

This document provides some supplementary material and additional results relating to Magnus and Wang (2013). It contains the within-group correlation of five groups in which we re-sign a variable, the BACE results (Sala-i-Martin et al., 2004) which we use to compare our results with, and the full results using data-dependent prior probabilities. It provides the procedure and detailed results of sensitivity analysis with respect to the prior π and to the grouping. It also contains a growth empirics study using another data set to test the robustness of the endogenous growth model.

6.2. Scaling

We present five groups in which we re-sign a variable in Table 6.1, so that variables within one group are positively correlated.

Note that the original variables in each group are highly and negatively correlated. Therefore, averaging estimates without scaling cancels the effect of these variables.

6.3. Results of BACE and HWALS-F1

The original results of BACE in Sala-i-Martin et al. (2004) are posterior mean and standard deviation estimates, conditional on inclusion along with posterior conditional probabilities. Since model uncertainty is not fully taken into account in the posterior standard deviations conditional on inclusion, the precision of the estimates is misleading as explained in Magnus et al. (2010). To 'fairly' compare the estimates produced by BACE with those of WALS and HWALS, we compute the unconditional ('true') moments

Table 6.1: Within-group correlations

g	Group	v	Variable	Correlation
(1)	Demographic characteristics	1*	Fraction population over 65	
		2	Fraction population under 15	−0.91
(2)	Economy system	3	Capitalism	−0.58
		4*	Socialist dummy	
(5)	Health	19	Life expectancy in 1960	−0.73
		20*	Malaria prevalence in 1960s	
(8)	Democracy	25	Political rights	−0.83
		26*	Civil liberties	
(11)	Tropics effect	31	Fraction of tropical area	−0.89
		32	Tropical climate zone	−0.60
		33*	Absolute latitude	

* Adjusted variable.

of BACE, based on Equations (8) and (14) in Sala-i-Martin et al. (2004).

The unconditional posterior mean can be computed by multiplying the conditional mean times the posterior inclusion probability, and the unconditional variance can be calculated as

$$\sigma^2_{uncond} = \left(\sigma^2_{cond} + \beta^2_{cond}\right) \times (posterior\ inclusion\ prob) - \beta^2_{uncond}.$$

Both the conditional and the unconditional estimates of BACE are given in Table 6.2, where the variables are ordered in the same way as in Magnus and Wang (2013).

Results of HWALS-F1 are also provided in Table 6.2 as a direct comparison between HWALS and BMA because *all* explanatory variables are allowed to be either included or excluded.

Table 6.2: BACE results and HWALS-F1

Variable	BACE results		HWALS-F1
	Conditional posterior estimates	Unconditional posterior estimates	
Education			−0.0013 (0.0038)
5 Primary schooling	0.0269 (0.0080)	0.0214 (0.0130)	
6 Secondary schooling			
7 Higher education	−0.0697 (0.0418)	−0.0043 (0.0196)	
8 Public edu. spending	0.1295 (0.1728)	0.0027 (0.0312)	
9 Primary school yrs			
10 Secondary school yrs			
11 Higher education yrs			
12 Ave. school yrs			
13 Ave. school yrs × logGDP			
Health			0.0045 (0.0044)
19 Life expectancy	0.0008 (0.0004)	0.0002 (0.0004)	
20 Malaria prevalence	−0.0157 (0.0062)	−0.0040 (0.0075)	
Initial state			−0.0030 (0.0046)
23 GDP in 1960 (log)	−0.0085 (0.0029)	−0.0058 (0.0046)	
24 Size of economy	−0.0005 (0.0014)	−0.0000 (0.0002)	
Tropics effect			−0.0029 (0.0032)
31 Frac. of tropical area	−0.0148 (0.0042)	−0.0083 (0.0080)	
32 Tropical climate zone	−0.0021 (0.0066)	0.0000 (0.0009)	
33 Absolute latitude	0.0001 (0.0002)	0.0000 (0.0000)	
Ethnicity and language			
36 Ethnolinguistic frac.	−0.0113 (0.0058)	−0.0012 (0.0039)	−0.0024 (0.0025)
37 English-speaking pop.	−0.0037 (0.0071)	−0.0001 (0.0011)	
38 Frac. foreign language	0.0070 (0.0040)	0.0006 (0.0022)	
Religion			
39 Fraction Confucian	0.0544 (0.0224)	0.0112 (0.0242)	
40 Fraction Muslim	0.0126 (0.0063)	0.0014 (0.0045)	
41 Fraction Buddhist	0.0217 (0.0107)	0.0023 (0.0076)	
42 Fraction Protestant	−0.0119 (0.0093)	−0.0005 (0.0032)	
43 Fraction Hindu	0.0176 (0.0126)	0.0008 (0.0045)	
44 Fraction Catholic	−0.0084 (0.0085)	−0.0003 (0.0022)	
45 Fraction Orthodox	0.0057 (0.0136)	0.0001 (0.0018)	
46 Religious intensity	−0.0047 (0.0072)	−0.0001 (0.0012)	−0.0009 (0.0017)
Price distortion			
70 Investment price	−0.0001 (0.0000)	−0.0001 (0.0000)	−0.0029 (0.0015)

Table 6.2: Continued

Variable	BACE results		HWALS-F1
	Conditional posterior estimates	Unconditional posterior estimates	
Demographic characteristics			0.0026 (0.0047)
1 Frac. pop. over 65	0.0194 (0.1195)	0.0004 (0.0180)	
2 Frac. pop. under 15	0.0450 (0.0411)	0.0018 (0.0122)	
Economy system			−0.0010 (0.0016)
3 Capitalism	−0.0002 (0.0011)	0.0000 (0.0001)	
4 Socialist dummy	0.0040 (0.0050)	0.0001 (0.0009)	
Government intervention			−0.0004 (0.0021)
14 Public investment share	−0.0615 (0.0430)	−0.0030 (0.0162)	
15 Public consumption share (excl. education and defense)			
16 Gov. consumption share	−0.0442 (0.0254)	−0.0046 (0.0158)	
17 Gov. share of GDP	−0.0349 (0.0294)	−0.0022 (0.0112)	
18 Nominal gov. GDP share	−0.0336 (0.0274)	−0.0012 (0.0081)	
Inflation			0.0005 (0.0022)
21 Average inflation	−0.0001 (0.0001)	−0.0000 (0.0000)	
22 Square of inflation	0.0000 (0.0000)	0.0000 (0.0000)	
Democracy			0.0025 (0.0027)
25 Political rights	−0.0018 (0.0102)	−0.0001 (0.0005)	
26 Civil liberties	−0.0072 (0.0071)	−0.0002 (0.0017)	
Scale effect			0.0028 (0.0027)
27 Land area	0.0000 (0.0000)	0.0000 (0.0000)	
28 Population	0.0000 (0.0000)	0.0000 (0.0000)	
Trade policy indices			0.0010 (0.0025)
29 Outward orientation	−0.0033 (0.0027)	−0.0001 (0.0007)	
30 Years open	0.0122 (0.0063)	0.0015 (0.0045)	
War			−0.0001 (0.0017)
34 Frac. spent in war	−0.0014 (0.0092)	−0.0000 (0.0012)	
35 War participation	−0.0007 (0.0030)	−0.0000 (0.0004)	
Trade statistics			
47 Openness measure	0.0089 (0.0052)	0.0007 (0.0028)	0.0007 (0.0029)
48 Primary exports	−0.0113 (0.0075)	−0.0006 (0.0031)	
Terms of trade			
49 Terms of trade ranking	−0.0037 (0.0096)	−0.0001 (0.0013)	0.0004 (0.0027)
50 Terms of trade growth	0.0326 (0.0467)	0.0007 (0.0082)	0.0036 (0.0024)
Regional effect			
51 East Asian dummy	0.0218 (0.0061)	0.0179 (0.0100)	0.0062 (0.0028)
52 African dummy	−0.0147 (0.0069)	−0.0023 (0.0060)	−0.0033 (0.0035)
53 European dummy	−0.0023 (0.0105)	−0.0001 (0.0019)	0.0016 (0.0045)
54 Latin-American dummy	−0.0128 (0.0058)	−0.0019 (0.0051)	−0.0012 (0.0045)
55 Colony dummy	−0.0050 (0.0047)	−0.0001 (0.0012)	−0.0040 (0.0035)
56 British colony	0.0037 (0.0036)	0.0001 (0.0008)	0.0028 (0.0027)
57 Spanish colony	−0.0107 (0.0050)	−0.0013 (0.0039)	0.0015 (0.0033)

Table 6.2: Continued

Natural resource			
58 Hydrocarbon deposits	0.0003 (0.0004)	0.0000 (0.0001)	0.0001 (0.0019)
59 Frac. GDP in mining	0.0388 (0.0193)	0.0048 (0.0145)	−0.0014 (0.0019)
60 Oil country dummy	0.0048 (0.0071)	0.0001 (0.0012)	−0.0018 (0.0023)
Population			
61 Population density coastal	0.0000 (0.0000)	0.0000 (0.0000)	0.0010 (0.0029)
62 Interior density	0.0000 (0.0000)	0.0000 (0.0000)	−0.0011 (0.0017)
63 Fraction pop. in tropics	−0.0107 (0.0068)	−0.0006 (0.0030)	0.0014 (0.0032)
64 Population density	0.0000 (0.0000)	0.0000 (0.0000)	−0.0016 (0.0021)
65 Population growth rate	0.0208 (0.3078)	0.0004 (0.0425)	0.0013 (0.0053)
66 Fertility	−0.0075 (0.0101)	−0.0002 (0.0022)	−0.0031 (0.0061)
Geography (excl. tropics effect)			
67 Frac. land area near water	−0.0026 (0.0059)	0.0000 (0.0009)	0.0018 (0.0032)
68 Landlocked country dummy	−0.0021 (0.0042)	0.0000 (0.0007)	0.0003 (0.0018)
69 Air distance to big cities	0.0000 (0.0000)	0.0000 (0.0000)	0.0010 (0.0025)
Real exchange rate			
71 Real exchange rate dist.	−0.0001 (0.0000)	0.0000 (0.0000)	−0.0024 (0.0020)
Defense			
72 Defense spending share	0.0453 (0.0768)	0.0010 (0.0129)	−0.0003 (0.0017)
Political instability			
73 Revolutions and coups	−0.0071 (0.0061)	−0.0002 (0.0016)	−0.0005 (0.0019)
Independence			
74 Timing of independence	0.0011 (0.0021)	0.0000 (0.0003)	0.0006 (0.0025)

6.4. Complete results of data-dependent prior

In Section 2.3 of the paper we discussed two updating algorithms based on data-dependent priors: one-step updating and two-step updating. This section provides complete results of HWALS-F8 using two updating methods.

In Table 6.4 and 6.5 we present the updated priors and the new HWALS-F8 estimates for all groups. The robustness of the updated probabilities and resulting estimates confirms that model specification only has a marginal effect in the updating procedure.

Table 6.4: HWALS estimates using data-dependent priors: focus variables

Variable	One-step updating HWALS-F8	updated π	Two-step updating HWALS-F8	updated π
Education	0.0051 (0.0034)		0.0050 (0.0034)	
5 Primary schooling		0.9784		0.9784
6 Secondary schooling		0.0033		0.0033
7 Higher education		0.0024		0.0024
8 Public edu. spending		0.0027		0.0027
9 Primary school yrs		0.0025		0.0025
10 Secondary school yrs		0.0031		0.0031
11 Higher education yrs		0.0019		0.0019
12 Ave. school yrs		0.0024		0.0024
13 Ave. school yrs × logGDP		0.0033		0.0033
Health	0.0062 (0.0059)		0.0065 (0.0060)	
19 Life expectancy		0.8142		0.8142
20 Malaria prevalence		0.1858		0.1858
Initial state	−0.0084 (0.0057)		−0.0082 (0.0059)	
23 GDP in 1960 (log)		0.6923		0.6923
24 Size of economy		0.3077		0.3077
Tropics effect	−0.0041 (0.0034)		−0.0040 (0.0034)	
31 Frac. of tropical area		0.5488		0.5488
32 Tropical climate zone		0.1489		0.1489
33 Absolute latitude		0.3023		0.3023
Ethnicity and language				
36 Ethnolinguistic frac.	−0.0022 (0.0026)	–	−0.0023 (0.0028)	–
37 English-speaking pop.		–		–
38 Frac. foreign language		–		–
Religion				
39 Fraction Confucian		–		–
40 Fraction Muslim		–		–
41 Fraction Buddhist		–		–
42 Fraction Protestant		–		–
43 Fraction Hindu		–		–
44 Fraction Catholic		–		–
45 Fraction Orthodox		–		–
46 Religious intensity	−0.0022 (0.0018)	–	−0.0022 (0.0018)	–
Price distortion				
70 Investment price	−0.0046 (0.0015)	–	−0.0045 (0.0016)	–

Table 6.5: HWALS estimates using data-dependent priors: auxiliary variables

	One-step updating		Two-step updating	
Variable	HWALS-F8	updated π	HWALS-F8	updated π
Demographic characteristics	0.0026 (0.0044)		0.0019 (0.0041)	
1 Frac. pop. over 65		0.7202		0.3282
2 Frac. pop. under 15		0.2798		0.6718
Economy system	−0.0007 (0.0015)		−0.0007 (0.0014)	
3 Capitalism		0.5227		0.5516
4 Socialist dummy		0.4773		0.4484
Government intervention	0.0005 (0.0020)		0.0004 (0.0020)	
14 Public investment share		0.1555		0.1647
15 Public consumption share (excl. education and defense)		0.1710		0.0220
16 Gov. consumption share		0.4317		0.5210
17 Gov. share of GDP		0.0827		0.0440
18 Nominal gov. GDP share		0.1590		0.2484
Inflation	0.0004 (0.0019)		0.0007 (0.0018)	
21 Average inflation		0.5005		0.5372
22 Square of inflation		0.4995		0.4628
Democracy	0.0015 (0.0024)		0.0017 (0.0024)	
25 Political rights		0.3017		0.2217
26 Civil liberties		0.6983		0.7783
Scale effect	0.0018 (0.0024)		0.0012 (0.0019)	
27 Land area		0.5293		0.1411
28 Population		0.4707		0.8589
Trade policy indices	0.0009 (0.0026)		0.0011 (0.0027)	
29 Outward orientation		0.0002		0.0005
30 Years open		0.9998		0.9995
War	−0.0003 (0.0015)		−0.0005 (0.0015)	
34 Frac. spent in war		0.5932		0.6160
35 War participation		0.4068		0.3840
Trade statistics				
47 Openness measure	−0.0003 (0.0025)	–	−1.50e-5(0.0025)	–
48 Primary exports		–		–
Terms of trade				
49 Terms of trade ranking	0.0002 (0.0024)	–	−0.0003 (0.0023)	–
50 Terms of trade growth	0.0021 (0.0022)	–	0.0019 (0.0022)	–
Regional effect				
51 East Asian dummy	0.0046 (0.0025)	–	0.0046 (0.0025)	–
52 African dummy	−0.0020 (0.0032)	–	−0.0018 (0.0032)	–
53 European dummy	0.0009 (0.0040)	–	0.0014 (0.0040)	–
54 Latin-American dummy	−0.0002 (0.0042)	–	0.0004 (0.0042)	–
55 Colony dummy	−0.0040 (0.0031)	–	−0.0036 (0.0030)	–
56 British colony	0.0022 (0.0026)	–	0.0021 (0.0025)	–
57 Spanish colony	0.0007 (0.0031)	–	0.0002 (0.0029)	–

Table 6.5: Continued

Variable	One-step updating HWALS-F8	One-step updated π	Two-step updating HWALS-F8	Two-step updated π
Natural resource				
58 Hydrocarbon deposits	0.0005 (0.0017)	–	0.0005 (0.0017)	–
59 Frac. GDP in mining	−0.0012 (0.0017)	–	−0.0012 (0.0017)	–
60 Oil country dummy	−0.0004 (0.0021)	–	−0.0003 (0.0020)	–
Population				
61 Population density coastal	0.0026 (0.0025)	–	0.0024 (0.0026)	–
62 Interior density	−0.0008 (0.0015)	–	−0.0007 (0.0015)	–
63 Fraction pop. in tropics	0.0009 (0.0028)	–	0.0007 (0.0028)	–
64 Population density	−0.0009 (0.0018)	–	−0.0007 (0.0017)	–
65 Population growth rate	0.0003 (0.0047)	–	0.0012 (0.0047)	–
66 Fertility	−0.0006 (0.0052)	–	−0.0004 (0.0052)	–
Geography (excl. tropics effect)				
67 Frac. land area near water	0.0001 (0.0030)	–	−0.0004 (0.0027)	–
68 Landlocked country dummy	−0.0003 (0.0016)	–	−0.0005 (0.0016)	–
69 Air distance to big cities	−0.0001 (0.0023)	–	−0.0003 (0.0022)	–
Real exchange rate				
71 Real exchange rate dist.	−0.0019 (0.0019)	–	−0.0019 (0.0019)	–
Defense				
72 Defense spending share	−0.0007 (0.0016)	–	−0.0008 (0.0016)	–
Political instability				
73 Revolutions and coups	−0.0003 (0.0017)	–	−0.0005 (0.0017)	–
Independence				
74 Timing of independence	0.0010 (0.0023)	–	0.0010 (0.0022)	–

6.5. Sensitivity with respect to the prior π

In Section 3.3 of the paper we distinguished between four levels of belief regarding the specification of the prior probabilities π. In Section 5 of the paper we present the empirical results based on the first two levels (default equal priors and data-dependent priors). In this section we investigate the effects of the belief of π on the estimates and standard deviations. In other words, we ask how sensitive the empirical results and the conclusions are with respect to π. We discuss the two levels in turn: equal priors and ordered priors. The analysis of equal priors corresponds to the first case where researchers have no information, and we study how the change in the prior of one variable affects the results, keeping priors of other variables equal; The other case of order priors corresponds to the third and fourth cases where researchers have unequal information and are able to order the variables in one group. The way of assigning priors and the effect of priors are very different from Ley and Steel (2009) or Eicher et al. (2011) who found sensitive results with respect to priors, because the priors in our case are assigned to the variables in one group, unlike the priors of the models or parameters. Therefore, we expect the effect of priors on our results is also different from the findings in Ley and Steel (2009) and Eicher et al. (2011).

6.5.1. Equal priors

Suppose that in one group, say group l, the m_l variables do not all have the same prior probability $1/m_l$, but that one of the variables, say variable j, has a different probability π_l^j, while the remaining variables have equal probabilities

$$\pi_l^i = \frac{1-\pi_l^j}{m_l - 1} \qquad (i = 1, \ldots, j-1, j+1, \ldots, m_l). \tag{6.1}$$

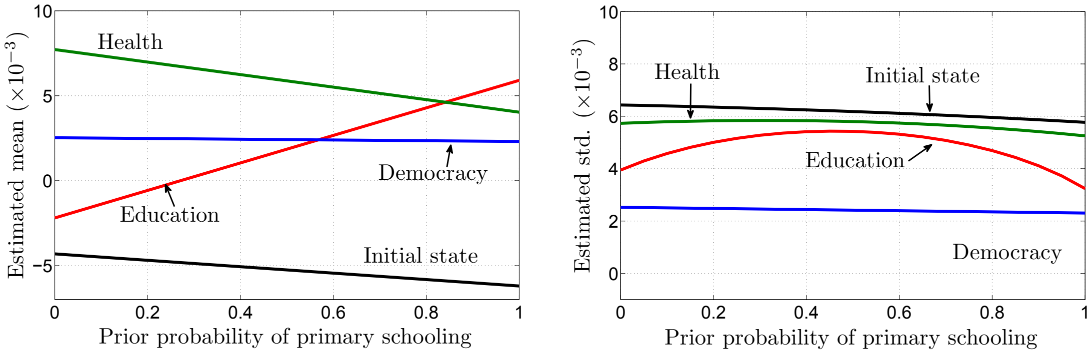
This assumption can be made for any of the type I groups. For our sensitivity experiment we choose one focus variable 'education' and one auxiliary variable 'democracy'. We choose education, because this group contains nine variables with relatively large deviations, and the effect of education on economic growth depends on which variable is used; see Table 3 of the paper. Democracy is of interest because it has strong policy implications (Barro, 1999) and its effect on economic growth is controversial. Within education we choose the variable 'primary schooling' as the one whose prior probability is different, because its role in explaining economic growth appears to differ from other education variables; within democracy we choose the variable 'political rights'.

In Figure 6.1 we report the sensitivity of four groups: education, democracy, health, and initial state. These groups are chosen because they are proximate determinants that are typically regarded as the most important growth theories. Also, by including education and democracy, we can investigate the direct effect (effect on the group itself) and the indirect effect (effect on other groups) of changing the prior probability. In the figure, the prior probability of primary schooling (top panel) and political rights (bottom panel) varies between 0 and 1. Note that the estimated means of all variables are linear with the varying prior probability, because π_l^i in Equation (6.1) is a linear function of π_l^j.

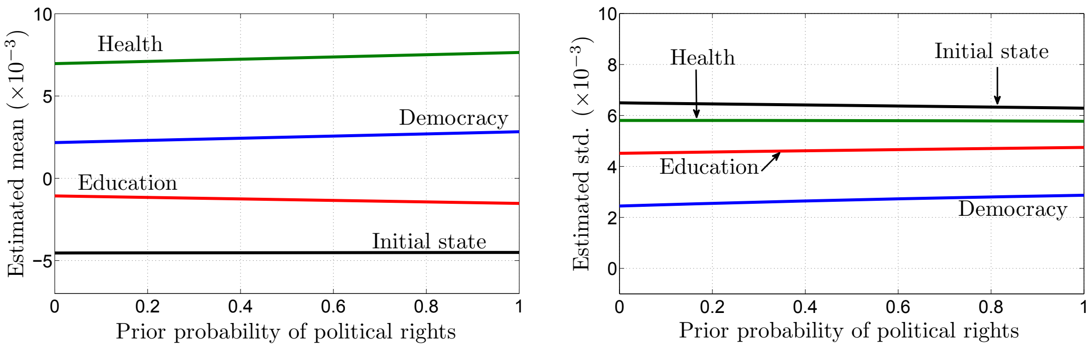
Changing the prior probability of primary schooling has a serious direct effect on the estimated mean and standard deviation of education. The estimated mean of education is negative when the prior probability of primary schooling is less than 0.2, but becomes positive when it is larger than 0.4. This is due to the fact that primary schooling has a strong positive effect on growth while the effect of other education variables is weak.

Figure 6.1: Sensitivity with respect to π: unequal priors

Unequal prior in group ‘education’



Unequal prior in group ‘democracy’



The estimated standard deviation of education is a concave function of the prior probability. It is obvious that when primary schooling has weight 1, we obtain the minimal standard deviation because there is no variation between variables. It is less obvious that when primary schooling has a small weight, we also obtain small standard deviations. The reason lies in the fact that primary schooling differs much from other education variables, so that a small weight leads to a small cross term bb' in Equation (12) of the paper.

The estimates in the other three groups (indirect effect) are less sensitive than those of education. The estimated means never change sign. Among the three groups, health appears to be the most sensitive to the change in prior. The estimated standard deviations of the three groups are all insensitive to the change in prior.

In the bottom panel we change the prior on political rights in the group democracy. The effects are very small. Even the estimate of democracy itself (the direct effect) is not sensitive.

6.5.2. Ordered priors

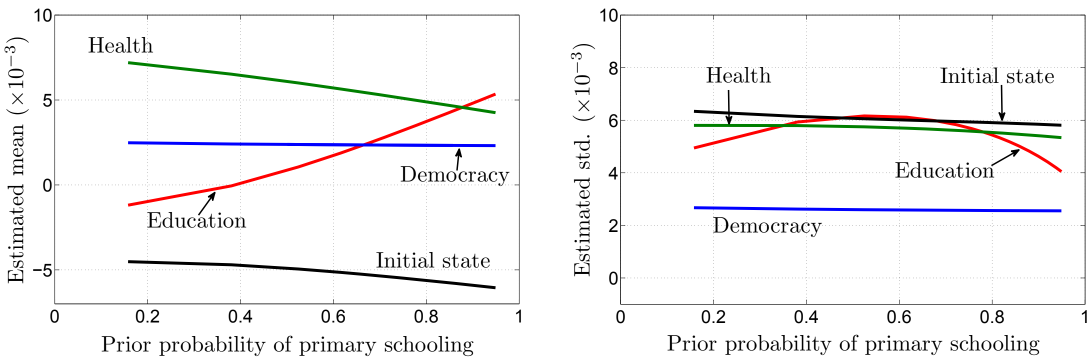
Next suppose that we can order the priors in group l so that the priors of the variables $x_l^1, \ldots, x_l^{m_l}$ are known to satisfy $\pi_l^1 > \cdots > \pi_l^{m_l}$. In particular, assume that $\pi_l^{i+1} = r\pi_l^i$ for some $0 < r < 1$. Then,

$$\pi_l^i = \frac{(1-r)r^{i-1}}{1-r^{m_l}} \qquad (i = 1, \ldots, m_l). \tag{6.2}$$

The smaller is r the more weight is placed on the important variables. Equation (6.2) allows the prior probability of the most important measurement, π_l^1, to change over the interval $(1/m_l, 1)$.

Figure 6.2: Sensitivity with respect to π: ordered priors

Ordered priors in group ‘education’



Ordered priors in group ‘democracy’

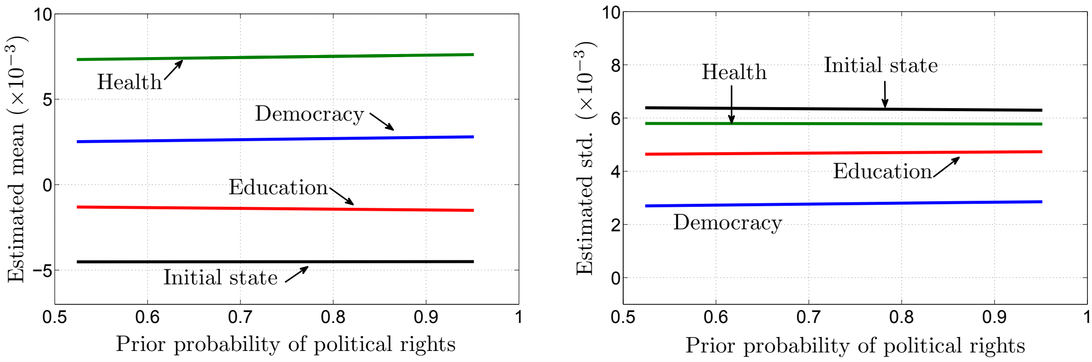


Figure 6.2 presents some representative examples when the priors are ordered. To perform this experiment we set $r = 1/2$ and we need a predetermined ordering of the variables. In the group education we select primary schooling as the most important variable and we order the other variables randomly. Unlike the previous case, neither the estimated mean nor the standard deviation is a linear function of the prior probability.

Still, the main results are essentially the same as before. In particular, as the prior probability of primary schooling increases, the estimated mean of education changes from negative to positive, and the estimated standard deviation of education is a concave function. The health effect is weakened as the prior probability of primary schooling increases, while initial state and democracy are insensitive to the probability change.

In the group democracy we select political rights as the most important variable. In this group there are only two variables, so that $\pi^1 = 1/(1+r)$ and $\pi^2 = r/(1+r)$. Changing the probability hardly changes the estimated means and standard deviations for any of the groups.

We repeated these experiments for all other type I groups, both for the unequal prior case and for the ordered priors case. Based on these experiments we draw three conclusions regarding the sensitivity with respect to the prior probability. First, the effects of proximate determinants on economic growth is robust to the choice of prior probability, except for the group education. Second, the indirect effects of the prior probability are very small, while the direct effect varies across groups. The direct effect is large for those groups whose variables vary greatly, such as education. But it is small for those groups whose variables are highly correlated, such as health, inflation, and scale effect. Third, the standard deviations are quite robust.

6.6. Sensitivity with respect to grouping

It is not always easy to decide which variable belongs in which group. In this section we investigate the sensitivity of estimates and standard deviations with respect to grouping. We consider six scenarios. First, we consider separating GDP per capita in 1960 and the initial size of the economy. This is motivated by the neoclassical growth model where initial GDP per capita has a structural role and thus should always be included. This is scenario S1. Second, we question whether the variable 'public spending on education' belongs in the group 'education'. This variable has a low correlation with other education variables, so perhaps it should be placed in a separate group (scenario S2). Third, one might make a case for placing this variable in the group 'government intervention' (scenario S3). Besides separating public education spending, one could also consider the possibility that enrollment rates and attainment levels (school years) have a different effect on growth, because the former is a flow measure while the latter captures the stock of human capital. Thus, in scenario S4, we separate enrollment rates (variables 5–7 and 8), school years (variables 9–13), and public education spending. In scenario S5, we allow that lower level (primary and secondary) and higher education may have different effects because the first is related with basic literacy necessary for simpler activities while the latter provides advanced capability useful in some innovative industry. Finally, we consider separating latitude from tropic effect group since it could also measure proximity to major economic hubs (scenario S6).

Table 6.6 presents the results of focus variables under alternative groupings. We see that S1 leads to a much larger estimated coefficient ($b = -0.0098$) of the initial level of income and a smaller variance ($V = 0.0053$), making initial income as one of the most important determinants explaining cross-country growth differences. This also provides strong evidence of convergence. The new grouping also has an impact on other focus variables, but this impact is not large. For example, education effect on steady-state growth is still weakly negative, while the effects of health, ethnolinguistic fractionalization, and religion are strengthened to different extents. Results of other focus variables and auxiliary variables are marginally affected except that estimated coefficients of democracy and scale effects reduced by around 28% and 32%, respectively, and that of trade statistic is doubled (standard deviation unchanged). Separating public education spending makes education group slightly more negative but with larger variance given by column S2, while assigning public education spending in 'government intervention' group (column S3) has a weaker effect on education than S2. Column S4 shows that the effects of both the flow measure and the stock measure of education are weak. Distinguishing different levels of education (S5) suggests that primary and secondary education has a very weak effect on growth, while the effect higher education is even negative. These results confirm large variation of education variables as well as their distinct effects on growth. In general we see most education variables are weakly related with growth except for primary schooling. Finally, we see from column S6 that it does not make much difference distinguishing between absolute latitude and other variables in the tropic effect group.

Table 6.6: Sensitivity analysis on grouping: focus variables

	S1	S2	S3	S4	S5	S6
Education	−0.0002 (0.0043)	−0.0021 (0.0050)	−0.0017 (0.0048)			−0.0015 (0.0048)
Edu.stock				−0.0020 (0.0066)		
Edu.flow				−0.0011 (0.0042)		
Lower edu.					0.0000 (0.0049)	
Higher edu.					−0.0044 (0.0032)	
Health	0.0088 (0.0064)	0.0077 (0.0060)	0.0074 (0.0058)	0.0077 (0.0060)	0.0082 (0.0062)	0.0071 (0.0058)
Initial state	−0.0098 (0.0053)	−0.0046 (0.0065)	−0.0045 (0.0064)	−0.0043 (0.0062)	−0.0042 (0.0063)	−0.0048 (0.0064)
Tropics effect	−0.0030 (0.0033)	−0.0029 (0.0034)	−0.0031 (0.0034)	−0.0032 (0.0034)	−0.0031 (0.0034)	−0.0028 (0.0041)
Ethnicity and lang.	−0.0035 (0.0028)	−0.0028 (0.0028)	−0.0030 (0.0028)	−0.0024 (0.0028)	−0.0024 (0.0028)	−0.0030 (0.0028)
Religion	−0.0017 (0.0019)	−0.0017 (0.0019)	−0.0015 (0.0019)	−0.0018 (0.0019)	−0.0017 (0.0020)	−0.0016 (0.0020)
Price distortion	−0.0041 (0.0017)	−0.0042 (0.0017)	−0.0041 (0.0017)	−0.0044 (0.0017)	−0.0045 (0.0017)	−0.0040 (0.0018)

Notes: Public education spending is treated as auxiliary in S2–S5, and other education-related groups are focus.

We also look at the average relative change between the HWALS-F8 estimates and standard deviations from the new grouping and the original grouping. If we consider all variables, then some changes in grouping have a relatively significant effect on the estimates, but the standard deviations are not much affected. This applies in particular to the case where public spending in education is separated from group 'education', and we find 30% change in estimates and 0.01% change in standard deviations. Grouping public spending into government intervention affects the results only marginally, with 6% change in estimates and 0.6% standard deviation change. Distinguishing between lower level and higher education (S4) and flow and stock (S5) lead to moderate changes in both estimates and standard deviations, and separating absolute latitude also has a weak effect. The average relative change is much less if we only compare the focus groups.

6.7. Growth models: Alternative data set

This section applies hierarchical model averaging to a small set of growth data. There are three reasons for doing this. First, we test the robustness of the endogenous growth model with the distinction between focus and auxiliary regressors. Second, we compare our results with Magnus et al.'s (2010) Model 2 that uses a similar data set. Finally, the small data set allows us to study *all* combinations of measurements, so that we can have more information on the distribution of our estimates, not just the first two moments.

This data set introduces two variables that are not listed in Tables 1 and 2 of our main paper, namely

75 Equipment investment

76 Rule of law index,

both taken from Sala-i-Martin (1997). We follow a similar specification as in Model 2 of Magnus et al. (2010), in which nine focus variables (labeled 'F' in Table 6.7) and three auxiliary regressors were considered. The auxiliary variables are: political rights (25), fraction GDP in mining (59), and population growth rate (65).

We note two deviations from the specification in Magnus et al. (2010). First, we take into account different alternative measurements of four type I groups, namely education, health, democracy, and tropics effect. Second, since malaria prevalence belongs to the group 'health', it is not estimated as a separate auxiliary variable as in Magnus et al. (2010).

Table 6.7: Results for small data set

Group/Variable	WALS	HWALS
Constant (F)	0.0211 (0.0013)	0.0211 (0.0014)
Type I groups/variables		
Education (F)		−0.0009 (0.0038)
5 Primary schooling (F)	0.0039 (0.0026)	
Health (F)		0.0060 (0.0045)
19 Life expectancy (F)	0.0065 (0.0041)	
20 Malaria prevalence	0.0022 (0.0018)	
Initial state (F)		−0.0062 (0.0068)
23 GDP in 1960 (log) (F)	−0.0149 (0.0030)	
Democracy		0.0008 (0.0027)
25 Political rights	−0.0014 (0.0018)	
Tropics effect (F)		0.0005 (0.0024)
31 Frac. of tropical area (F)	−0.0017 (0.0019)	
Type II variables		
36 Ethnolinguistic frac. (F)	−0.0020 (0.0018)	−0.0026 (0.0022)
39 Frac. Confucian (F)	0.0049 (0.0015)	0.0060 (0.0019)
59 Frac. GDP in mining	−0.0003 (0.0013)	−0.0006 (0.0015)
65 Population growth rate	0.0015 (0.0021)	0.0013 (0.0025)
75 Equipment investment (F)	0.0041 (0.0020)	0.0053 (0.0024)
76 Rule of law index (F)	0.0074 (0.0024)	0.0073 (0.0031)

The estimates based on the small data set are given in Table 6.7. The WALS results in the table are different in magnitude from Magnus et al. (2010) (with the same signs) due to the scaling of the regressors and also the different number of observations (We have 6 countries less than Magnus et al. (2010) since we include additional alternative measurements.) For type I groups, the WALS estimates correspond to variables while the estimates in HWALS correspond to groups.

We first compare the signs, and then the precisions as in the large data set. Three groups have different signs: education, tropics effect, and democracy. The counterintuitive sign of education produced by HWALS is mainly due to large variation of nine measurements as discussed in our main paper. For the group tropics effect, HWALS produces a positive estimate but very insignificant. The standard deviation is more than four times the mean, so that the sign of the mean estimate is very uncertain. The large standard deviation is mainly caused by the fact that the variables 'tropic climate zone' and 'absolute latitude' are insignificant and not robust. If we fix this group to 'fraction of tropical area' and re-estimate using HWALS, *ceteris paribus*, we obtain a strongly negative tropics effect. As for the group democracy, HWALS reports a negative effect, but insignificant as well. This is in line with most studies on the association between growth and democracy, as discussed in more detail in our paper.

Next, we comment on the precision of the estimates. Unlike in the large data set, estimates from HWALS, in this case, have larger standard deviations than estimates from WALS. This is because HWALS standard deviation explicitly takes into account additional uncertainty on the choice of measurement that is not considered in WALS. It is also because there are not many type I groups in this small data set, and thus not many highly correlated variables are included in the WALS regressions. Therefore, multicollinearity is not a serious problem here. To gain further insight on how the esti-

Table 6.8: Quantiles of estimated mean

Quantile	0.1	0.3	0.5	0.7	0.9
Constant (F)	0.0211	0.0211	0.0211	0.0211	0.0211
Education (F)	−0.0044	−0.0022	−0.0012	0.0005	0.0033
Health (F)	0.0029	0.0037	0.0054	0.0066	0.0118
Initial state (F)	−0.0138	−0.0119	−0.0054	−0.0001	0.0004
Democracy	−0.0016	−0.0002	0.0007	0.0017	0.0036
Tropics effect (F)	−0.0013	−0.0003	0.0009	0.0015	0.0019
Ethnolinguistic frac. (F)	−0.0038	−0.0033	−0.0026	−0.0020	−0.0013
Frac. Confucian (F)	0.0048	0.0052	0.0058	0.0068	0.0073
Frac. GDP in mining	−0.0010	−0.0008	−0.0005	−0.0004	−0.0001
Population growth rate	−0.0001	0.0009	0.0015	0.0018	0.0024
Equipment investment (F)	0.0039	0.0051	0.0055	0.0058	0.0063
Rule of law index (F)	0.0052	0.0065	0.0073	0.0079	0.0096

mated mean and standard deviation vary across different combinations of measurements, we present the quantiles of the mean and standard deviation estimates in Tables 6.8 and 6.9. The quantiles are obtained based on 216 ($= 2 \times 9 \times 2 \times 3 \times 2$) possible choices of measurements. The quantiles are calculated without weighting, and therefore invariant

Table 6.9: Quantiles of estimated standard deviation

Quantile	0.1	0.3	0.5	0.7	0.9
Constant (F)	0.0013	0.0013	0.0013	0.0015	0.0015
Education (F)	0.0018	0.0020	0.0020	0.0028	0.0031
Health (F)	0.0021	0.0022	0.0022	0.0033	0.0036
Initial state (F)	0.0021	0.0022	0.0022	0.0032	0.0034
Democracy	0.0017	0.0018	0.0018	0.0020	0.0021
Tropics effect (F)	0.0015	0.0017	0.0017	0.0023	0.0025
Ethnolinguistic frac. (F)	0.0018	0.0020	0.0020	0.0021	0.0021
Frac. Confucian (F)	0.0015	0.0016	0.0016	0.0016	0.0017
Frac. GDP in mining	0.0013	0.0013	0.0013	0.0016	0.0016
Population growth rate	0.0020	0.0022	0.0022	0.0024	0.0026
Equipment investment (F)	0.0020	0.0021	0.0021	0.0023	0.0024
Rule of law index (F)	0.0024	0.0025	0.0025	0.0029	0.0029

to the prior probability. Quantiles provide more information on the distribution of the estimates, from which we can see which groups (variables) are robust with respect to the choice of measurements and which are not. We emphasize that it makes no sense to associate the estimated mean and standard deviation at the same quantile level because they are obtained from different measurement combinations.

We first look at the quantiles of the estimated means in Table 6.8. For type I groups, health is the most robust group with positive mean estimates at all quantiles. Although initial state, democracy, and tropics effect are less robust, still about 70% of their estimated means are consistently negative. Education has the largest variation in mean estimates. About half of them are negative, while the other half are positive. The estimated means of type II groups all have consistent signs except population growth rate. Among them, the variables equipment investment, religion, and rule of law are highly robust, while ethnolinguistic fraction, fraction GDP in mining, and population growth rate are relatively less robust. Generally speaking, we observe that estimates of variables (type II groups) vary much less than estimates of concepts (type I groups). This suggests that changing the measurements of a group has a much smaller impact on other groups (indirect effect) than on the group itself (direct effect). This observation is the basis of our approximation strategy. Compared to the estimated means, the estimated standard deviations reported in Table 6.9 have much smaller variations, showing that estimated standard deviations are more robust to the choice of measurements than means.

Bibliography

Acemoglu, D., S. Johnson, and J. A. Robinson (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review 91*, 1369–1401.

Aiolfi, M. and A. Timmermann (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics 135*, 31–53.

Auty, R. M. (2001). The political economy of resource-driven growth. *European Economic Review 45*, 839–846.

Barro, R. J. (1996). Democracy and growth. *Journal of Economic Growth 1*, 1–27.

Barro, R. J. (1999). Determinants of democracy. *The Journal of Political Economy 107*, S158–S183.

Barro, R. J. and J.-W. Lee (1993). International comparisons of educational attainment. *Journal of Monetary Economics 32*, 363–394.

Bates, J. M. and C. W. J. Granger (1969). The combination of forecasts. *Operational Research Quarterly 20*, 451–468.

Beck, T. and R. Levine (2005). Legal institutions and financial development. In S. M. Ménard C (Ed.), *Handbook for New Institutional Economics*, pp. 251–278. Dordrecht: Kluwer Academic Publishers.

Belleflamme, P. and J. Hindriks (2005). Yardstick competition and political agency problems. *Social Choice and Welfare 24*, 155–169.

Beran, R. (1995). Prediction in random coefficient regression. *Journal of Statistical Planning and Inference 43*, 205–213.

Besley, T. and A. Case (1995). Incumbent behavior: Vote-seeking, tax-setting, and yardstick competition. *American Economic Review 85*, 25–45.

Besley, T. and M. Smart (2007). Fiscal restraints and voter welfare. *Journal of Public Economics 91*, 755–773.

Bjørnland, H. C., K. Gerdrup, A. S. Jore, C. Smith, and L. A. Thorsrud (2012). Does forecast combination improve norges bank inflation forecasts? *Oxford Bulletin of Economics and Statistics 74*, 163–179.

Bodenstein, M. and H. Ursprung (2005). Political yardstick competition, economic integration, and constitutional choice in a federation. *Public Choice 124*, 329–352.

Bondeson, J. (1990). Prediction in random coefficient regression models. *Biometrical Journal 32*, 387–405.

Bordignon, M., F. Cerniglia, and F. Revelli (2004). Yardstick competition in intergovernmental relationships: Theory and empirical predictions. *Economics Letters 83*, 325–333.

Brock, W. A. and S. N. Durlauf (2001). Growth empirics and reality. *World Bank Economic Review 15*, 229–272.

Brock, W. A., S. N. Durlauf, and K. D. West (2003). Policy evaluation in uncertain economic environments. *Brookings Papers on Economic Activity*, 235–301.

Brunnschweiler, C. N. and E. H. Bulte (2008). The resource curse revisited and revised: A tale of paradoxes and red herrings. *Journal of Environmental Economics and Management 55*(3), 248–264.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997). Model selection: An integral part of inference. *Biometrics 53*, 603–618.

Bulte, E. H., R. Damania, and R. T. Deacon (2005). Resource intensity, institutions, and development. *World Development 33*, 1029–1044.

Cai, H. and D. Treisman (2005). Does competition for capital discipline governments? Decentralization, globalization, and public policy. *American Economic Review 95*, 817–830.

Cai, Z., J. Fan, and Q. Yao (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association 95*, 941–956.

Chipman, H. (1996). Bayesian variable selection with related predictors. *The Canadian Journal of Statistics 24*, 17–36.

Ciccone, A. and M. Jarociński (2010). Determinants of economic growth: Will data tell? *American Economic Journal: Macroeconomics 2*, 222–246.

Claessens, S. and L. Laeven (2003). Financial development, property rights, and growth. *Journal of Finance 58*, 2401–2436.

Collier, P. and B. Goderis (2008). Commodity prices, growth, and the natural resource curse: Reconciling a conundrum. MPRA Paper 17315, University Library of Munich, Germany.

Danilov, D. (2005). Estimation of the mean of a univariate normal distribution when the variance is not known. *Econometrics Journal 8*, 277–291.

Danilov, D. and J. R. Magnus (2004a). Forecast accuracy after pretesting with an application to the stock market. *Journal of Forecasting 23*, 251–274.

Danilov, D. and J. R. Magnus (2004b). On the harm that ignoring pretesting can cause. *Journal of Econometrics 122*, 27–46.

De Luca, G. and J. R. Magnus (2011). Bayesian model averaging and weighted average least squares: Equivariance, stability, and numerical issues. *The Stata Journal 11*, 518–544.

Doppelhofer, G. and M. Weeks (2009). Jointness of growth determinants. *Journal of Applied Econometrics 24*, 209–244.

Durlauf, S. N., P. A. Johnson, and J. R. W. Temple (2005). Growth econometrics. In S. N. Durlauf (Ed.), *Handbook of Economic Growth*, pp. 555–677. Amsterdam: North Holland.

Durlauf, S. N., A. Kourtellos, and C. M. Tan (2008). Are any growth theories robust? *The Economic Journal 118*, 329–346.

Durlauf, S. N., A. Kourtellos, and C. M. Tan (2012). Is God in the details? A reexamination of the role of religion in economic growth. *Journal of Applied Econometrics 27*, 1059–1075.

Eicher, T. S., C. Papageorgiou, and A. E. Raftery (2011). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics 26*, 30–55.

Elliott, G. and A. Timmermann (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics 122*, 47–79.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. London, UK: Chapman and Hall.

Fan, R., Y. Fang, and S. Y. Park (2012). Resource abundance and economic growth in China. *China Economic Review*.

Fernandez, C., E. Ley, and M. F. J. Steel (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics 16*, 563–576.

Froehlich, B. R. (1973). Some estimators for a random coefficient regression model. *Journal of the American Statistical Association 68*, 329–335.

George, E. I. (2010). Dilution priors: Compensating for model space redundancy. In J. O. Berger, T. T. Cai, and I. M. Johnstone (Eds.), *Borrowing Strength: Theory Powering Applications — A Festschrift for Lawrence D. Brown*, pp. 158–165. Beachwood, Ohio, USA: Institute of Mathematical Statistics.

Giles, J. A. and D. E. A. Giles (1993). Pre-test estimation and testing in econometrics: Recent developments. *Journal of Economic Surveys 7*, 145–197.

Golub, G. H., M. Heath, and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics 21*, 215–223.

Granger, C. W. J. (2003). Time series concepts for conditional distributions. *Oxford Bulletin of Economics and Statistics 65*, 689–701.

Habakkuk, H. J. (1962). *American and British Technology in the Nineteenth Century*. Cambridge, MA: Cambridge University Press.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*, 1175–1189.

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics 146*, 342–350.

Hendry, D. F. and M. P. Clements (2004). Pooling of forecasts. *Econometrics Journal 7*, 1–31.

Hildreth, C. and J. P. Houck (1968). Some estimators for a linear model with random coefficients. *Journal of the American Statistical Association 63*, 584–595.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*, 879–899.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Hoover, D. R., J. A. Rice, C. O. Wu, and L. Yang (1998). Nonparametric smoothing estimates of time-varying coeffcient models with longitudinal data. *Biometrika 85*(4), 809–822.

Huang, J., J. L. Horowitz, , and F. Wei (2010). Variable selection in nonparametric additive models. *The Annals of Statistics 38*, 2282–2313.

Isham, J., M. Woolcock, L. Pritchett, and G. Busby (2005). The varieties of resource experience: Natural resource export structures and the political economy of economic growth. *World Bank Economic Review 19*, 141–174.

Jackson, T. and S. Karlsson (2004). Finding good predictors for inflation: A Bayesian model averaging approach. *Journal of Forecasting 23*, 479–496.

Janeba, E. and W. Peters (1999). Tax evasion, tax competition and the gains from nondiscrimination: The case of interest taxation in Europe. *Economic Journal 109*, 93–101.

Jensen, P. S. and A. H. Würtz (2012). Estimating the effect of a variable in a high-dimensional linear model. *Econometrics Journal 15*, 325–357.

Johnston, J. and J. DiNardo (1997). *Econometric Methods* (Fourth ed.). New York: McGraw-Hill.

Kapetanios, G. (2008). A bootstrap procedure for panel datasets with many cross-sectional units. *Econometrics Journal 11*, 377–395.

Knack, S. and P. Keefer (1995). Institutions and economic performance: Cross-country tests using alternative institutional measures. *Economics and Politics 7*, 207–227.

Leamer, E. E. (1985). Sensitivity analyses would help. *American Economic Review 75*, 308–13.

Leamer, E. E. and H. B. Leonard (1983). Reporting the fragility of regression estimates. *The Review of Economics and Statistics 65*, 306–317.

Leeb, H. and B. M. Pötscher (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory 19*, 100–142.

Leeb, H. and B. M. Pötscher (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics 34*, 2554–2591.

Leeb, H. and B. M. Pötscher (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *The Annals of Statistics 24*, 338–376.

Leite, C. and J. Weidmann (1999). Does mother nature corrupt - natural resources, corruption, and economic growth. IMF working papers, International Monetary Fund.

Levitt, S. D. (2002). Using electoral cycles in police hiring to estimate the effects of police on crime: Reply. *American Economic Review 92*, 1244–1250.

Ley, E. and M. F. Steel (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics 24*, 651–674.

Liang, H., G. Zou, A. T. K. Wan, and X. Zhang (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association 106*, 1053–1066.

Lockwood, B. (2005). Fiscal decentralization: A political economy perspective. Warwick Economic Research Papers 721, University of Warwick.

Magnus, J. R. (1999). The traditional pretest estimator. *Theory of probability and its applications 44*, 293–308.

Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal 5*, 225–236.

Magnus, J. R., O. Powell, and P. Prüfer (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics 154*, 139–153.

Magnus, J. R., A. T. K. Wan, and X. Zhang (2011). Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Computational Statistics and Data Analysis 55*, 1331–1341.

Magnus, J. R. and W. Wang (2013). Concept-based Bayesian model averaging and growth empirics. Discussion paper.

Mani, A. and S. Mukand (2007). Democracy, visibility and public good provision. *Journal of Development Economics 83*, 506–529.

Mankiw, N. G., D. Romer, and D. N. Weil (1992). A contribution to the empirics of economic growth. *The Quarterly Journal of Economics 107*, 407–37.

Mehlum, H., K. Moene, and R. Torvik (2006). Institutions and the resource curse. *Economic Journal 116*, 1–20.

Ministry of Land and Resources (2000-2009). *China land and resources statistical yearbook* (2000-200 ed.). Beijing: Geological Publishing House.

National Bureau of Statistics (2003-2004). *China statistical yearbook* (2003-2004 ed.). Beijing: China Statistics Press.

Norman, C. (2009). Rule of law and the resource curse: Abundance versus intensity. *Environmental & Resource Economics 43*, 183–207.

Nunn, N. (2007). Relationship-specificity, incomplete contracts, and the pattern of trade. *The Quarterly Journal of Economics 122*, 569–600.

Ohtani, K. and T. Toyoda (1980). Estimation of regression coefficients after a preliminary test for homoscedasticity. *Journal of Econometrics 12*, 151–159.

Olken, B. A. (2009). Corruption perceptions vs. corruption reality. *Journal of Public Economics 93*, 950–964.

Papyrakis, E. and R. Gerlagh (2004). The resource curse hypothesis and its transmission channels. *Journal of Comparative Economics 32*, 181–193.

Poirier, D. J. (1995). *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, MA: MIT Press.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*, 179–191.

Rogoff, K. (1990). Equilibrium political budget cycles. *American Economic Review 80*, 21–36.

Ross, M. L. (2001). *Timber Booms and Institutional Breakdown in Southeast Asia*. Cambridge, MA: Cambridge University Press.

Rubin, H. (1950). Note on random coefficients. In T. C. Koopmans (Ed.), *Statistical Inference in Dynamic Economic Models*, pp. 419–421. Cowles Commission Monograph No. 10.

Ruud, P. A. (2000). *An Introduction to Classical Econometric Theory*. New York/Oxford: Oxford University Press.

Sachs, J. D. and A. M. Warner (1995). Natural resource abundance and economic growth. NBER Working Papers 5398, National Bureau of Economic Research, Inc.

Sala-i-Martin, X. (1997). I just ran two million regressions. *American Economic Review 87*, 178–183.

Sala-i-Martin, X., G. Doppelhofer, and R. I. Miller (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review 94*, 813–835.

Salimans, T. (2012). Variable selection and functional form uncertainty in cross-country growth regressions. *Journal of Econometrics 171*, 267–280.

Shao, S. and Z. Qi (2009). Energy exploitation and economic growth in western China: An empirical analysis based on the resource curse hypothesis. *Frontiers of Economics in China 4*, 125–152.

Stock, J. H. and M. W. Watson (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting 23*, 405–430.

Sun, Y., R. J. Carroll, and D. Li (2009). Semiparametric estimation of fixed-effects panel data varying coefficient models. In Q. Li and J. S. Racine (Eds.), *Nonparametric Econometric Methods (Advances in Econometrics, Volume 25)*, pp. 101–129. Emerald Group Publishing Limited.

Swamy, P. A. V. B. (1970). Efficient inference in a random coefficient regression model. *Econometrica 38*, 311–23.

The World Bank (1999). *World Development Report 1999/2000: Entering the 21st Century*. New York: Oxford University Press.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting, Vol. 1*, pp. 135–196. Amsterdam: Elsevier.

Toyoda, T. and T. D. Wallace (1976). Optimal critical values for pre-testing in regression. *Econometrica 44*, 365–75.

van der Ploeg, F. and S. Poelhekke (2010). The pungent smell of "red herrings": Subsoil assets, rents, volatility and the resource curse. *Journal of Environmental Economics and Management 60*, 44–55.

Walley, P. and T. L. Fine (1982). Towards a frequentist theory of upper and lower probability. *Annals of Statistics 10*, 741–761.

Wan, A. T. K., X. Zhang, and G. Zou (2010). Least squares model averaging by mallows criterion. *Journal of Econometrics 156*, 277–283.

Wan, A. T. K. and G. Zou (2003). Optimal critical values of pre-tests when estimating the regression error variance: analytical findings under a general loss structure. *Journal of Econometrics 114*, 165–196.

Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning 24*, 149–170.

Whittle, P. (1963). *Prediction and Regulation by Linear Least-Square Methods.* London: The English Universities Press Ltd.

Xu, K. and J. Wang (2006). An empirical study of a linkage between natural resource abundance and economic development. *Economic Research Journal*, 78–89.

Yang, Y. (2004). Combining forecasting procedures: Some theoretical results. *Econometric Theory 20*, 176–222.

Zhang, X., L. Xing, S. Fan, and X. Luo (2008). Resource abundance and regional development in China. *The Economics of Transition 16*, 7–29.

Zissimos, B. and M. Wooders (2008). Public good differentiation and the intensity of tax competition. *Journal of Public Economics 92*, 1105–1121.